NAIST-IS-DD1461002

Doctoral Dissertation

Statistical waveform modification for speaking and singing voice conversion

Kazuhiro Kobayashi

March 16, 2017

Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Kazuhiro Kobayashi

Thesis Committee:

Prof. Satoshi Nakamura	(Supervisor)
Prof. Kenji Sugimoto	(Co-supervisor)
Prof. Tomoki Toda	(Nagoya University)
Assist. Prof. Sakriani Sakti	(Co-supervisor)
Dr. Masataka Goto	(AIST)

Statistical waveform modification for speaking and singing voice conversion *

Kazuhiro Kobayashi

Abstract

The variation of the voice characteristics, such as voice timbre and fundamental frequency (F_0) patterns, produced by individual speakers/singers is always restricted by their own physical constraints due to the speech production mechanism. In particular, the voice timbre of speaking and singing voice significantly depends on the physical constraints of the individual speakers/singers. If the speakers/singers freely produced many varieties of the voice timbre beyond their own physical constraints, it would break down the limitations of speaking and singing voice expressions and open up an entirely new speaking/singing expression style. In this thesis, toward the realization of a new speaking/singing expression style, we focus on two kinds of approaches for a user (speaker and singer) as follows: 1) convert the individuality of the user's voice expression into that of a specific target speaker/signer, and 2) control the voice factor of the source voice based on words expressing voice timbre as a perceptually understandable voice timbre definition.

Voice conversion (VC) is a potential technique for enabling us to produce speech sounds beyond our own physical constraints. As one of the most popular statistical VC methods, a regressian mixture model (GMM) was proposed. In this technique, the individuality of a source speaker/singer is converted into that of a target speaker/singer by altering several acoustic features such as F_0, aperiodicity, and spectral envelopes. However, the VC based on the GMM method is not used in practice because the sound quality of the converted voice is significantly degraded compared with that of a natural speech waveform.

One of the major factors causing the quality degradation is the waveform generation process using a vocoder. In the waveform generation process, a converted voice is generated by vocoding using transformed F_0, converted aperiodicity, and spectral envelopes. In this process, various factors such as F_0 extraction errors, unvoiced/voiced decision errors, and spectral parameterization errors resulting from liftering usually

[*] Doctoral Dissertation, Graduate School of Information Science,
Nara Institute of Science and Technology, NAIST-IS-DD1461002, March 16, 2017.

cause the sound quality degradation of the converted voice. To address this issue, in this thesis, we propose a statistical waveform modification technique for both VC and voice factor control approaches using speaking and singing voices.

For VC, to alleviate the sound quality degradation caused by vocoding, we propose a conversion technique without vocoding that directly modifies the signal in the waveform domain by estimating the difference between the spectra of the source and target voices. To convert the voice timbre of the source speaker/singer into that of the target speaker/singer, the spectral differential is estimated using a differential GMM (DIFFGMM), which was modeled by parameter transformation of the traditional GMM. Then, the converted voice is generated by filtering the waveform signal of the source voice with the estimated spectral differential. In this framework, it is possible to avoid the waveform generation process using vocoding because source excitation of the source voice is directly used as the excitation signal. The experimental results demonstrate that intra/inter-gender statistical waveform modification techniques make it possible to significantly improve the sound quality of the converted voice in VC and singing VC (SVC).

For voice factor control, we focus on the singing voice because singers have more opportunities to control their singing voice characteristics compared with speakers, because the singers often use vocal effectors. We focus on the perceived age, that is, the age that a listener predicts the singer to be, of singing voices as one of the factors to intuitively describe the singing voice. To fully understand the acoustic features that contribute to the perceived age of singing voices, we first perform an investigation of the acoustic features that play a part in the listener's perception of the singer's age. Then, we propose voice timbre control techniques based on multiple-regression GMM (MR-GMM), which converts the singer's perceived age while retaining singer individuality in SVC. The experimental results indicate that the proposed voice timbre control based on the perceived age makes it possible to control voice timbre while retaining singer identity.

Finally, we propose several techniques for implementing real-time VC and voice timbre control systems based on statistical waveform modification. We describe overviews and components of real-time VC and voice timbre control systems with statistical waveform modification. The experimental results demonstrate that our proposed real-time conversion systems make it possible to achieve higher sound quality than that obtained with the conventional real-time conversion framework.

Keywords:

statistical waveform modification, voice conversion, voice timbre control, speaker and singer individuality, Gaussian mixture model

Contents

List of Figures

List of Tables

1. Introduction

1.1. General background

In speech communication, speakers can handle their speech production mechanism subconsciously to transmit information. The variation of the voice characteristics, such as voice timbre and fundamental frequency (F_0) patterns, produced by individual speakers is always restricted by their own physical constraints due to the speech production mechanism. These constraints are helpful for making it possible to produce a speech signal capable of simultaneously conveying not only linguistic information but also nonlinguistic information such as speaker identity and emotions. However, the physical constraints sometimes cause various limitations to the expansion of speech expressions. In particular, voice timbre significantly depends on the physical constraints of an individual speaker. Figure 1.1 indicates the limitations of the voice timbre expression in each speaker. For speaker A, although he can produce a speaking voice consisting of various voice timbre expressions, such as joy, anger, and sadness, within his expression range, restricted by his physical constraints, he cannot produce a speaking voice with the voice timbre of the other speakers such as speakers B and C, because the voice timbres of these speakers is beyond the range of speaker A.

Singers face similar limitations of voice timbre expression because singing voices also rely on their speech production mechanism. To make singing voices more expressive, singers usually use various vocal effectors such as chorus, flangers, and pitch correction. Although these effectors are effective for augmenting their vocal expressions, it is difficult to understand the usage of vocal effectors intuitively because the control knobs of these effectors are usually labeled with unpredictable cues such as tone, mix, and depth. Moreover, these vocal effectors cannot enable voice timbre control beyond the singer's individuality (e.g., changing the voice timbre of a user into that of another specific singer). If individual speakers/singers could freely produce various

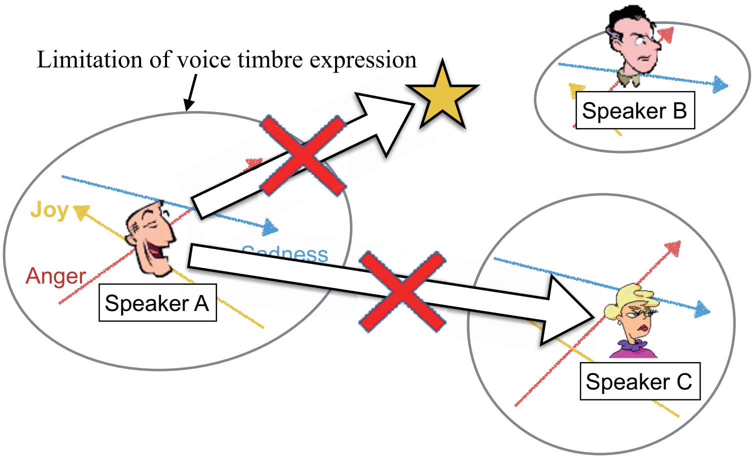


Figure 1.1.: Limitations of speaking expression for individual speakers.

voice timbre expressions beyond their physical constraint, it would break down the limitations and open up entirely new speech/singing expression styles.

To realize a new expression style, in this thesis, we focus on voice conversion and voice factor control to generate an ideal speaking/singing voice. Figure 1.2 illustrates our intended voice conversion and voice factor control systems. We assume that there are two different approaches to creating a desirable voice as follows: 1) convert the speaker/singer individuality of a user into that of another specific target speaker/signer, and 2) control the speaking/singing voice of a user in accordance with a perceptually understandable voice factor. The first approach is for the user to imitate the speaking/singing style of an ideal speaker/singer by assuming the voice characteristics of the ideal target speaker/singer. This can be achieved by converting the acoustic features of the user's speaking/singing voice into those of the ideal speaker/singer's speaking/singing voice using a voice conversion framework. In the latter approach, the user creates a new speaking/singing expression demonstrating his/her own creativity, without reference to a specific target speaking/singing voice. In this approach, to find an ideal speaking/singing voice intuitively, perceptually understandable factors such as voice timbre expression words (e.g., age, sweetness, and masculinity) are effective for controlling the speaking/singing voice.

In addition to implementing the above two approaches, we consider the following requisites for voice conversion and voice factor control systems to be important for

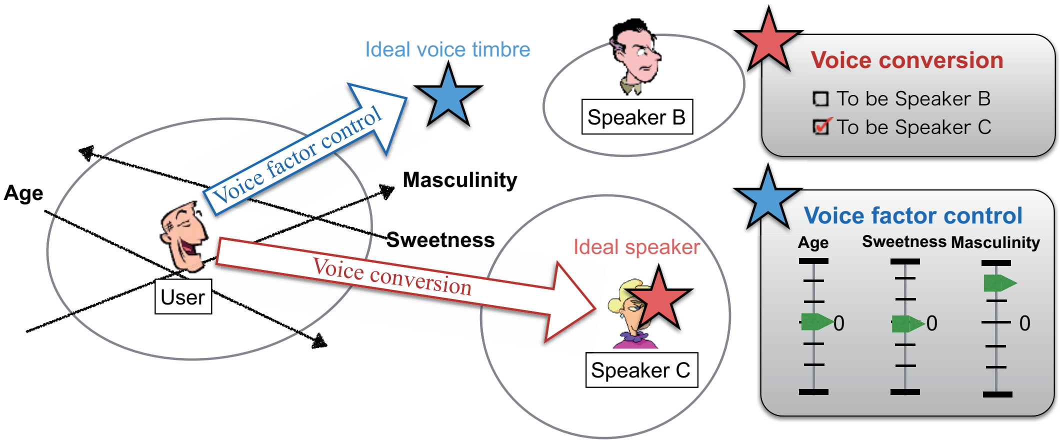


Figure 1.2.: Approaches of voice conversion and voice factor control.

satisfying the expectations of the user.

- Similarity: how similar the converted speaking/singing voice is to the ideal speaking/singing voice
- Controllability: how intuitively and correctly controllable the speaking/singing voice factor is
- Sound quality: the amount of sound quality degradation after conversion/control
- Real-time availability: conversion/control without any delay after inputting original speaking/singing voice.

When these requisites are satisfied, it is expected that the voice conversion and voice factor control systems can be used in practical applications.

1.2. Definition of voice conversion

Figure 1.3 illustrates the definition of voice conversion (VC). VC research was originally started as a technique for converting the individuality of a source speaker into that of a target speaker without changing the linguistic information of the source speech [1]. This conversion technique has been widely adapted into various research domains such as speech disorder [2], speech to articulatory mapping [3, 4], speech to facial image

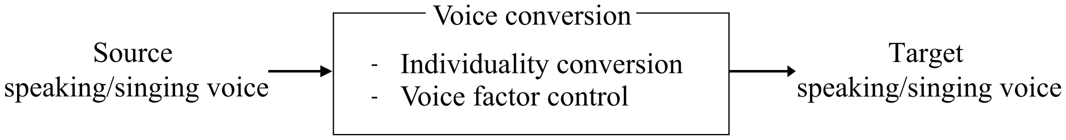


Figure 1.3.: Definition of voice conversion.

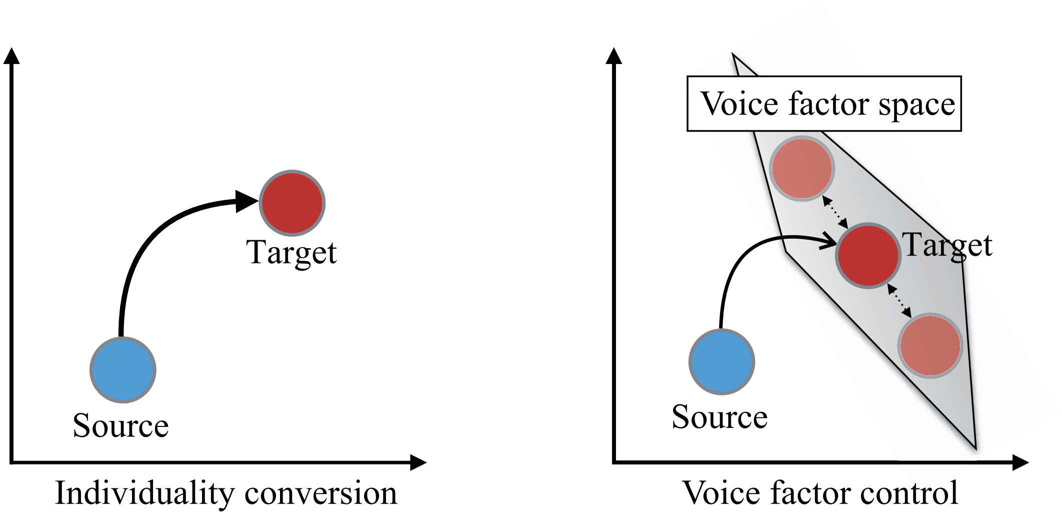


Figure 1.4.: Schematic image of voice conversion and voice factor control.

conversion [5], and speech enhancement [6]. Different from these domains, VC is usually called a speaker/singer individuality conversion technique. On the other hand, the applicability of VC is limited to speaker/singer individuality conversion and voice factor control. In this thesis, we separately deal with VC research on two different approaches, VC (i.e., speaker/singer individuality conversion) and voice factor control, because the main motivations behind these approaches are different. Figure 1.4 indicates a schematic image speaker/singer individuality conversion and voice factor control.

For VC, the research aim is to implement an absolute individuality conversion technique of speakers/singers when target speaking/singing voice samples are supplied to a system. In order to implement such a technique, all components of the source speaker/singer, such as acoustic features and locutions, are converted into those of the target speaker/singer. In other words, individuality conversion can be said to be a point estimation from the source speaker/singer the target speaker/singer.

In contrast to VC, the main aim of a voice factor control is to convert a source speaking/singing voice into an ideal speaking/singing voice without any specific target speaking/singing voice samples. To generate the ideal speaking/singing voice without the target, it is necessary to be capable of controlling the output voice because the system cannot assume the ideal voice timbre of the user. Therefore, the voice factor space, which takes on the controllability of voice timbre, should be not a point but a space. Moreover, the controlled components are chosen from among all components of the source speaking/singing voice because it is not necessary to control the components that do not affect the voice factor.

1.3. Related works and problem definition

VC [1] is a potential technique enabling VC system users to produce speaking/singing voices beyond their own physical constraints [7]. The mainstream in VC is a statistical approach to developing a conversion function using a parallel data set consisting of utterance pairs of the source and target speakers/singers. One of the most popular statistical VC methods is a regression method using a Gaussian mixture model (GMM) [8,9]. In this technique, acoustic features of the source speaker/singer are converted into those of the target speaker/singer on the basis of a previously trained GMM. The performance of the GMM-based VC method has been significantly improved by incorporating a trajectory-based conversion algorithm, modeling additional features to alleviate an oversmoothing effect of the converted speech parameters, such as global variance (GV) [10] and a modulation spectrum (MS) [11], and implementing sophisticated vocoding techniques such as STRAIGHT [12] with mixed excitation [13]. Moreover, the GMM-based VC method is capable of converting the voice timbre of an arbitrary source speaker/singer into that of an arbitrary target speaker/singer by the eigenvoice technique [14, 15], and controlling the voice timbre of the source speaker on the basis of perceptually understandable voice timbre expression words [16] such as age and powerful [17, 18]. In addition to these novel techniques for the GMM-based VC method, the most significant advantage compared with the other statistical conversion models such as Gaussian process regression [19,20] and deep neural networks [21–23] is that a small-delay conversion for a real-time VC system has been successfully implemented [24, 25].

Using the above novel techniques, among the four requisites defined in Section 1.1, “Similarity” and “Real-time availability” are almost satisfied compared with the other requisites, “Sound quality” and “Controllability”.

1.3.1. “Sound quality” problem

In VC based on GMM, its conversion process usually causes significant sound quality degradation of a converted voice owing to various factors such as insufficient modeling accuracies of the acoustic model and waveform generation. One of the major causes of this sound quality degradation is the waveform generation process using a vocoder [26]. For the vocoding process, the sound quality degradation is usually caused by various factors such as F_0 extraction errors, unvoiced/voiced decision errors, and spectral parameterization errors caused by liftering, which are indeed difficult to solve even when using high-quality vocoding frameworks [12, 27–29]. The waveform generation based on vocoding is usually a lossy process even when using the acoustic features extracted from an original waveform and without performing any modification. Therefore, the sound quality degradation is unavoidable as long as the vocoding framework is used.

1.3.2. “Controllability” problem

In VC for individuality conversion, it is not necessary to control the converted voice because the target is explicitly given by the user. On the other hand, for voice factor control, the controlling factor is defined on the basis of not an actual target but an ideal speaking/singing voice, and is constructed in the user’s mind. Therefore, it is difficult for a voice factor control system to estimate the control factor. In this thesis, to estimate the control factor and meet the expectations of the user, we focus on voice timbre expression words and their scores [17, 18]. Using the voice timbre expression words and their scores, it is expected that the system can tract the virtual target of the user.

As a technique of controlling the voice factor using voice timbre expression words, the voice timbre control technique based on the MR-GMM has been proposed [16]. This technique makes it possible to control the voice factor related to voice timbre in accordance with the voice timbre expression words. However, this technique can only convert the voice timbre of the source voice into an average voice corresponding to the voice timbre expression score input by the user. Therefore, it is impossible to maintain speaker/singer individuality after conversion.

1.4. Thesis scope

In this thesis, we focus on speaker/singer individuality conversion (i.e., VC) and singing voice factor control. At first, we describe a VC technique based on statistical waveform modification for a speaking/singing voice without the use of vocoding. Then, we describe a voice factor control technique based on perceptually understandable voice timbre expression words while retaining speaker/singer individuality. Finally, we describe our implemented real-time VC systems for both VC and voice factor control via statistical waveform modification.

1.4.1. Voice conversion

In Chapter 3, in order to alleviate sound quality degradation caused by vocoding, we deal with this issue in the order of intra-gender singing VC (SVC) and intra/inter-gender VC. Note that we rephrase the individuality conversion as VC following the convention in this research area.

Figure 1.5 illustrates the hypothesis of sound quality improvement without vocoding in VC. In Section 1.3.1, we described the problem of sound quality degradation when using the vocoding framework. For the conventional VC framework, it is impossible to avoid the use of the vocoding framework to transform F_0 and convert aperiodicity and spectral envelopes. It is expected that a vocoderless VC technique will make it possible to improve the sound quality of the converted voice.

Intra-gender SVC

First, we focus on intra-gender SVC. In contrast to VC for normal speech, in intra-gender SVC, it is not necessary to transform F_0 of the source singer into that of the target singer because these singers often sing in the same key defined by a song or

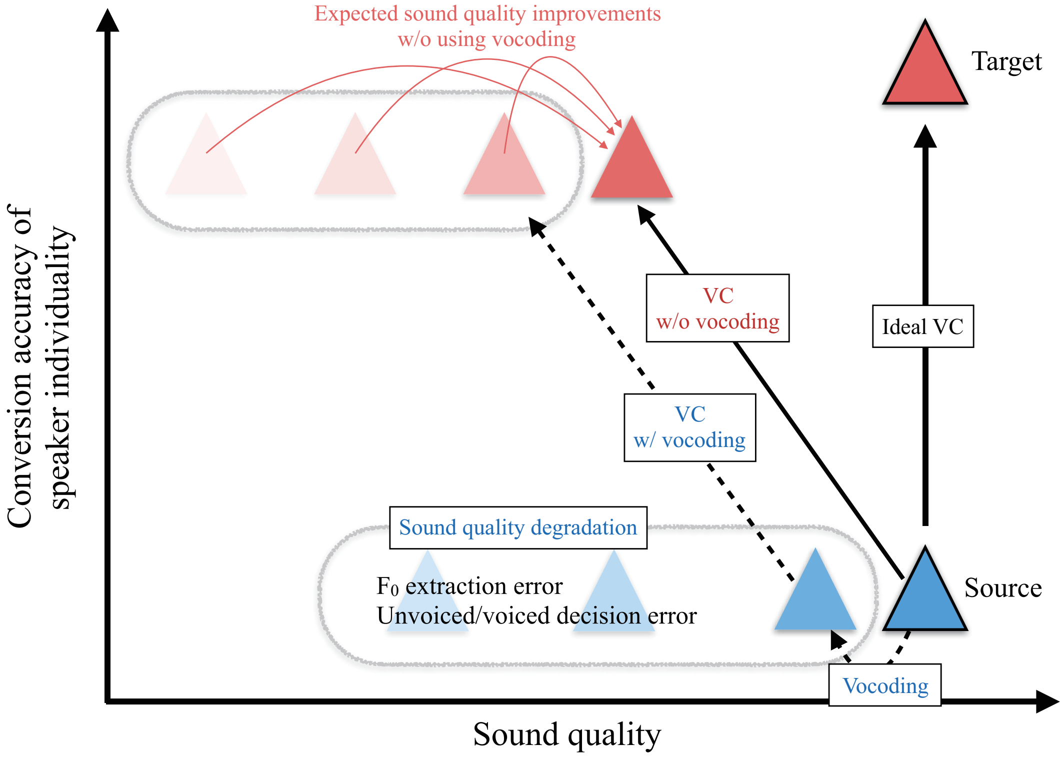


Figure 1.5.: Hypothesis of sound quality improvement without using vocoding.

score. Also, it is considered that the effect of the aperiodicity of intra-gender conversion does not exceed that of inter-gender conversion. Therefore, we assume that the conversion of spectral envelopes can achieve sufficient accuracy of the singer individuality conversion. On the basis of this idea, in intra-gender SVC, we propose the spectral conversion technique without the vocoder.

In the conventional VC framework, the use of the vocoding framework to generate a waveform of the converted voice is unavoidable because the source singing voice is fully decomposed into acoustic features to allow conversion into the target singing voice. This vocoding process causes significant sound quality-degradation of the converted voice. In Section 3.2, to avoid waveform generation using vocoding, we propose a statistical waveform modification technique of using the spectral differential. The waveform of the source singing voice is directly modified with a digital filter that uses the time-varying difference in the spectral envelope between the source and target singers' singing voices. Note that this spectrum differential is statistically estimated from the spectral envelopes of the source singing voice using a differential GMM (DIFFGMM). We propose several spectral differential generation techniques for intra-gender SVC using direct waveform modification (DIFFSVC): 1) spectral differential estimation based on the DIFFGMM, 2) spectral differential estimation based on the DIFFGMM considering GV, 3) spectral differential estimation based on the DIFFGMM considering MS, and 4) spectral differential estimation based on the trajectory differential spectral feature.

Intra/inter-gender VC

To make it possible to expand the intra-gender DIFFSVC framework to other situations such as inter-gender DIFFSVC or VC for normal speech, it is necessary to implement an F_0 transformation without vocoding because the F_0s of the source and target speakers/singers are usually different in such situations. In particular, the VC for normal speech is a more difficult problem compared with inter-gender SVC because it is necessary to accept not double or half but various F_0 transformation ratios. In Section 3.3, we propose several F_0 transformation techniques for intra/inter-gender VC based on direct waveform modification using the spectral differential (DIFFVC). The following F_0 transformation techniques with and without the vocoding process are proposed: 1) DIFFVC with F_0 transformation using a STRAIGHT vocoder, 2) DIFFVC with F_0 transformation based on the residual signal modification using time-scaling and resampling, and 3) DIFFVC with F_0 transformation based on waveform modification using time-scaling and resampling.

1.4.2. Voice factor control

In Chapter 4, we focus on the control of the singing voice timbre because the vocal effectors for the singing voice are widely used in practice compared with those of the speaking voice. Therefore, it is expected that a voice factor control system will be familiar to the user of singing vocal effectors.

For the voice factor control, we focus on voice timbre expression words [17] as control cues of the voice factor space. There are a great many voice timbre expression words such as sweet and powerful. It is considered that almost all voice timbre expression words will be effective in controlling the source voice if the user can use ideal voice timbre expression words. However, it is difficult to optimize the voice factor space for all users because annotated scores of voice timbre expression words tend to have no consistency with each user. Therefore, it is difficult to implement a generic voice factor control system for all users if the system accepts arbitrary voice timbre expression words.

In this thesis, to implement such a generic voice timbre control system for all users, we focus on the perceived age, that is the age that a listener predicts the singer to be, as one of the factors used to intuitively describe the singing voice. The age has several good properties: e.g., age is a measurement on the ratio scale unlike measurements on a nominal scale, such as gender; age is more understandable than other expressive word pairs because it is observable; the age is widely distributed over people. The perceived age is also expected to have some of these good properties and to be conveniently used as a control factor to continuously and intuitively modify singing voice characteristics. On the basis of this idea, we consider that the perceived age is appropriate as a voice timbre expression word in the first step of implementing the voice timbre control system.

In addition to implementing singing voice timbre control based on the perceived age, we also focus on the singer individuality. In the conventional voice timbre control based on the MR-GMM, the speaker individuality is one of the factors to be converted from the user to the target. Therefore, regardless of the difference of the users, the source voice individuality tends to be converted into another individuality. To address this issue, we propose voice timbre control based on the perceived age while retaining singer individuality.

In this thesis, first, we investigate the acoustic features affecting the perceived age and singer individuality. Then, we propose the voice timbre control technique based on the perceived age while retaining singer individuality.

Investigation of acoustic features affecting perceived age and singer individuality

There are several studies related to the age or the perceived age for normal speech. It has been reported that there is a correlation between the actual age and the perceived age [30]. As an investigation of the impact of aging on speech acoustics, it has been found that the aperiodicity of excitation signals tends to increase with age [31] and the perceived age for normal speech is varied by manipulating its F_0 variations, duration, and aperiodicity [32]. A method of classifying the speech of elderly and nonelderly people using spectral and prosodic features has also been developed [33]. On the other hand, the perceived age of singing voices has not been studied in detail yet. Therefore, it is not obvious whether these findings will also be found in singing voices.

A full understanding of the acoustic features that contribute to the perceived age of singing voices is essential to the development of VC techniques to modify a singer's perceived age. Therefore, in Section 4.2, we first perform an investigation of the acoustic features that play a part in the listener's perception of the singer's age. We conduct several types of perceptual evaluations to investigate 1) how well the perceived age of singing voices corresponds to the actual age of the singer, 2) whether or not SVC processing causes adverse effects on the perceived age of singing voices, 3) which spectral or prosodic features have a greater effect on the perceived age, and 4) which spectral or prosodic features include the individuality of a singer.

Voice timbre control based on perceived age while retaining singer identity

In Section 4.4, we propose a voice timbre control technique based on the perceived age while retaining singer individuality. Referring to the experimental results of the investigation described in Section 4.2, we indicated that both prosodic features (e.g., F_0 pattern) and spectral features have an effect on perceived age, and prosodic features more strongly affect the perceived age than spectral features but they also cause an adverse effect on the perceived singer's individuality. In the traditional SVC framework [9, 15], only the spectral features, such as the mel-cepstrum, are converted. In this section, we first apply VC based on MR-GMM [16] to SVC to achieve perceived age control. The standard MR-GMM has difficulty in maintaining the individuality of the source singer, because the subspace of the MR-GMM only expresses the average voice timbre of training singers. To solve this problem, we propose a voice timbre conversion technique with a modified MR-GMM to convert the singer's peceived age while retaining the singer's individuality. Moreover, towards the development of a better-controllability, higher quality, and more flexible framework than the voice timbre control based on MR-GMM, we also propose; 1) a method using gender-dependent MR-GMMs, 2) a method using direct waveform modification based on a spectrum differential, and 3) a rapid unsupervised adaptation method based on maximum a posteriori (MAP) estimation [34–36] to easily develop the singer-dependent MR-GMM.

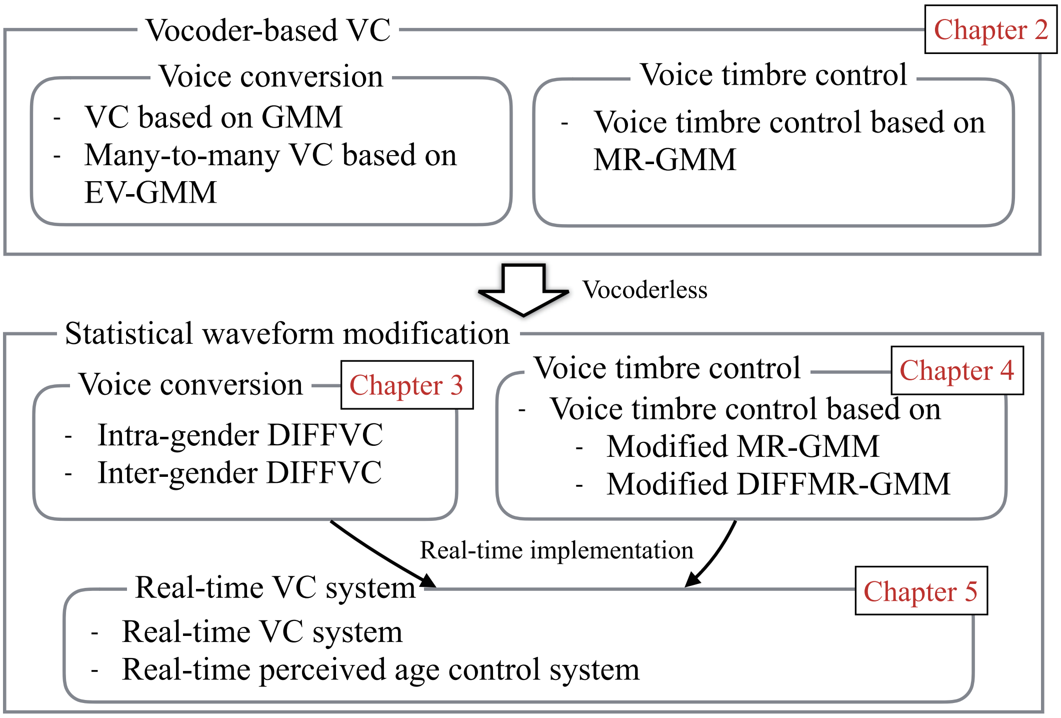


Figure 1.6.: Thesis overview.

1.4.3. Real-time VC systems based on statistical waveform modification

In Chapter 5, we describe our implemented real-time VC systems consisting of VC and voice timbre control via statistical waveform modification. It is not possible to directly apply a low-delay conversion algorithm [25] into a statistical waveform modification framework. To address this issue, we propose 1) a parameter transformation technique for the low-delay conversion algorithm, and 2) a frame-based GV postfilter for small-delay statistical waveform modification.

1.5. Thesis overview

Figure 1.6 shows an overview of this thesis. This thesis is organized as follows. In Chapter 2, vocoding and traditional VC frameworks are described as well as state-of-the-art conversion methods of those VC frameworks. In Chapter 3, we address the sound quality degradation caused by vocoding in intra-gender SVC and intra/inter-gender VC. First, in order to reduce the sound quality degradation of the converted voice, we propose intra-gender SVC based on direct waveform modification using the spectral differential (DIFFSVC). Then, to make it possible to apply intra-gender DIFFSVC to intra/inter-gender VC based on direct waveform modification using the spectral differential (DIFFVC), we propose several F_0 transformation techniques. The effectiveness of intra-gender DIFFSVC and intra/inter-gender DIFFVC is demonstrated by objective and subjective evaluations. In Chapter 4, we describe the voice timbre control technique for the singing voice based on perceived age while retaining singer identity. Lastly, we investigate acoustic features affecting the perceived age and singer individuality. Then, we propose a voice timbre control technique using the perceived age based on a modified representation of MR-GMM and differential MR-GMM (DIFFMR-GMM). Moreover, to easily use voice timbre control based on the modified DIFFMR-GMM, we propose an unsupervised adaptation technique of an arbitrary source speaker. The effectiveness of the voice timbre control technique based on the perceived age is confirmed by both objective and subjective evaluations. In Chapter 5, we describe our implemented real-time VC and voice timbre control systems and their components. The effectiveness of the real-time VC systems is confirmed by both objective and subjective evaluations. In Chapter 6, we summarize the contributions of this thesis and suggest future work.

2. Statistical voice conversion

2.1. Introduction

In this chapter, we describe following techniques related to statistical voice conversion (VC): 1) source-filter model of the speech production based on vocoding, 2) VC based on Gaussian mixture model (GMM), 3) many-to-many VC based on eigenvoice GMM (EV-GMM), 4) voice timbre control based on multiple-regression GMM (MR-GMM), and 5) a low-delay conversion algorithm for GMM-based VC.

This section is organized as follows: In Section 2.2, the source-filter model of speech production mechanism based on vocoding is illustrated. In Section 2.3, the framework of VC based on GMM is described. In Section 2.6, we explain many-to-many VC based on EV-GMM. In Section 2.7, we introduce statistical voice timbre control based on MR-GMM. In Section 2.8, we describe the low-delay conversion algorithm for GMM-based VC. In Section 2.9, we explain the issues of conventional VC frameworks. Finally, this chapter is summarized in Section 2.10.

2.2. Source-filter model of speech production based on vocoder

Figure 2.1 indicates a source-filter model of the speech production mechanism. Speech waveform has linguistic and non-linguistic information such as speaker/singer individuality, voice timbre, gender, words, and so on. In the source-filter model, the speech waveform is modeled as a combination of the source excitation signal and vocal tract parameter based on speech production mechanism of the individual speaker/singer. The source excitation signal is generated by vibrating vocal cord in regard to an exhaled breath from the lung. This source excitation signal plays an important role in prosody. In addition to add voice timbre derived from vocal tract parameter, speakers

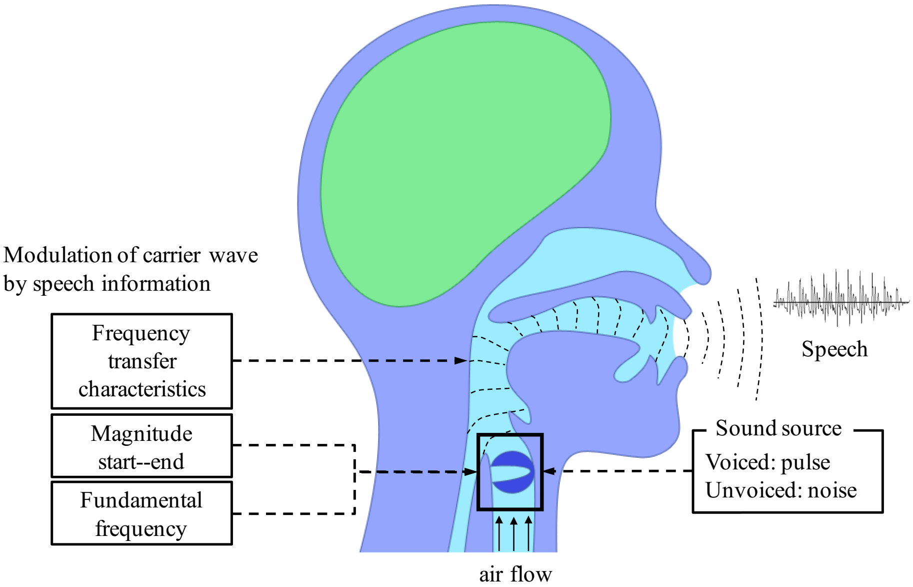


Figure 2.1.: Source-filter model based on speech production mechanism [37].

are capable of controlling the linguistic information by manipulating their articulator.

Figure 2.2 describes the analysis/synthesis process of the speech waveform based on vocoding. In the statistical parametric speech synthesis [38] including text-to-speech synthesis (TTS) and VC, F_0, aperiodic components and mel-cepstrum are widely hired as acoustic parameters in accordance with the source-filter model. For the analysis process, an original speech waveform is decomposed into F_0, aperiodic components, and mel-cepstrum based on STRAIGHT analysis [12]. For the synthesis process, at first, the source excitation signal is generated by the use of F_0 and aperiodic components in excitation generation. One-pitch signal is generated by selecting the phase-manipulated pulse signal based on the F_0, consisting of weighted noise signal using the aperiodic components, for voiced segments or white noise for an unvoiced segment. Then, pitch synchronous over-lap add (PSOLA) [39] is performed in order to generate the excitation signal. Finally, the speech waveform is generated by filtering the source excitation signal with the mel-cepstrum based on the mel log spectrum approximation (MLSA) filter [40].

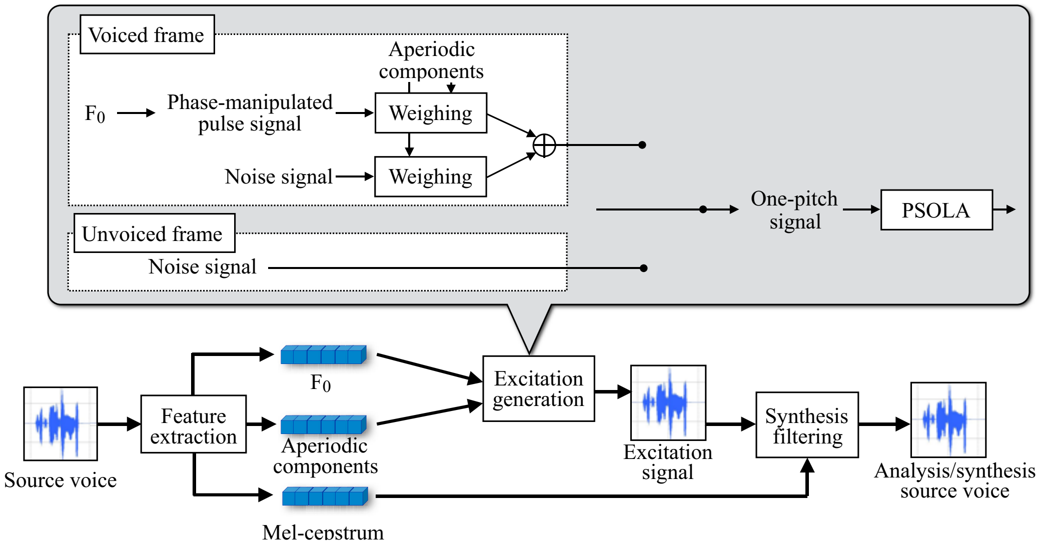


Figure 2.2.: Analysis/synthesis process based on vocoding.

2.3. VC based on GMM

VC based on GMM is a technique to convert voice timbre between a source speaker and a target speaker. This technique consists of the training process and conversion process. Figures 2.3 and 2.4 indicate the training and conversion process.

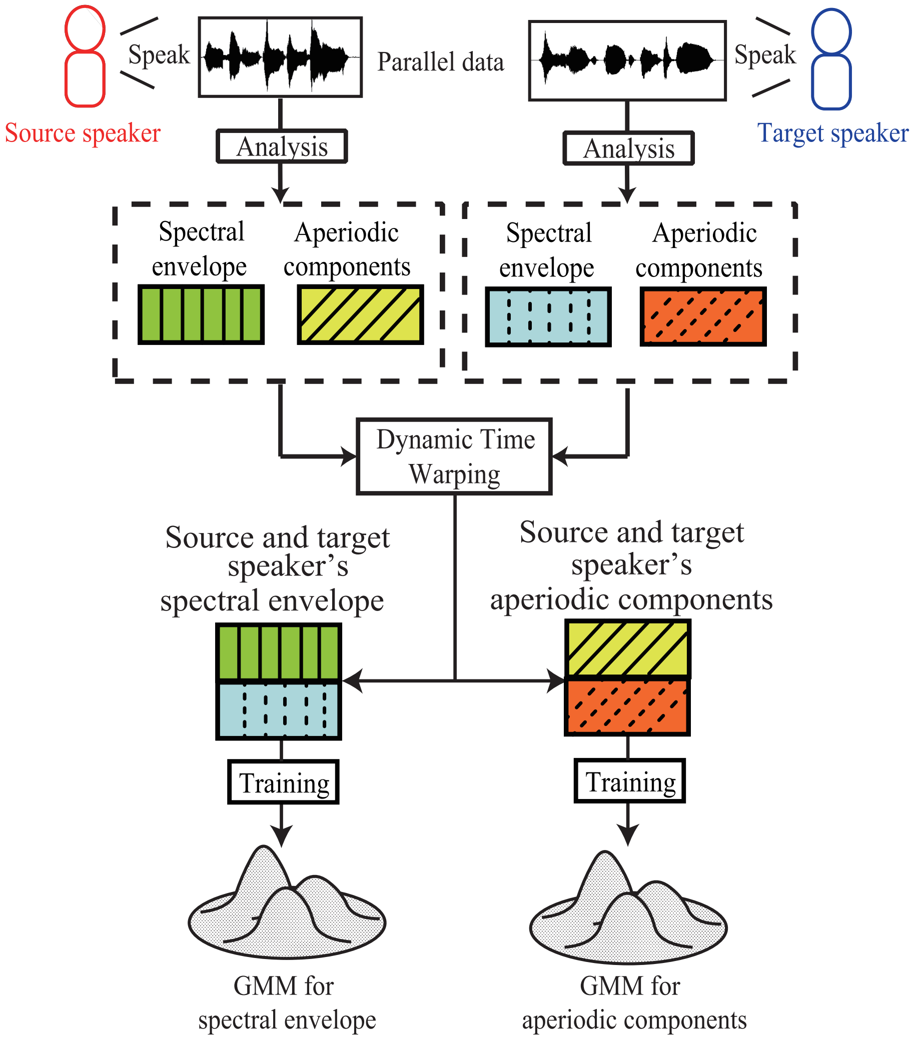


Figure 2.3.: Training process of VC based on GMM.

2.3.1. Training process of GMM

For the training process, a joint probability density function of acoustic features of the source and target speech samples is modeled with a GMM using a parallel data set [9]. As the acoustic features such as mel-cepstrum and aperiodic components of the source and target speakers, we employ $2D$-dimensional joint static and dynamic feature vectors $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$ of the source and $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$ of the target consisting of D-dimensional static feature vectors $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ and their dynamic feature vectors $\Delta\boldsymbol{x}_t$ and $\Delta\boldsymbol{y}_t$ at frame t, respectively, where $\top$ denotes the transposition of the vector. Their

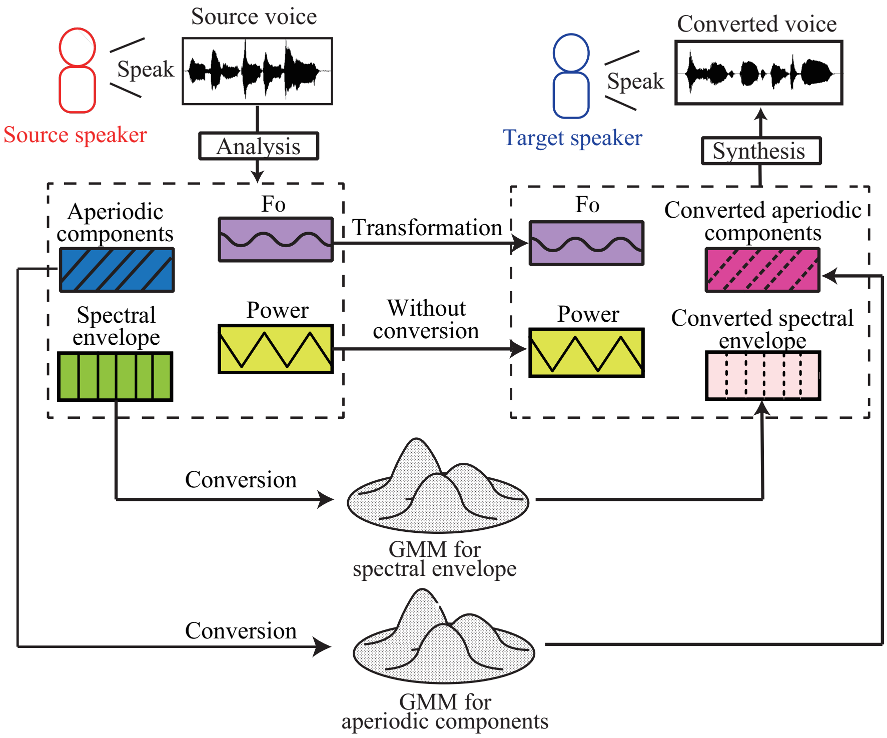


Figure 2.4.: Conversion process of VC based on GMM.

joint probability density modeled by the GMM is given by

$$P(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix} \boldsymbol{X}_t \\ \boldsymbol{Y}_t \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix}\right), \tag{2.1}$$

where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. The total number of mixture components is M. The mixture component index is m. $\boldsymbol{\lambda}$ is a GMM parameter set consisting of the mixture-component weight α_m, the mean vector $\boldsymbol{\mu}_m$, and the covariance matrix $\boldsymbol{\Sigma}_m$ of the m-th mixture component. A GMM is trained using joint vectors of $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ in the parallel data set, which are automatically aligned to each other by dynamic time warping.

2.3.2. Conversion process

For the conversion process, the acoustic feature extracted from a source speech sample is converted into target speaker based on maximum likelihood parameter generation technique [10]. Time sequence vectors of the source features and the target features are denoted as $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top]^\top$ and $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_T^\top]^\top$ where T is the number of frames included in the time sequence of the given source feature vectors.

$$\begin{aligned} P(\boldsymbol{Y}|\boldsymbol{X}, \lambda) &= \sum_{\text{all}\,\boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{X}, \lambda) P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{m}, \lambda) \\ &= \prod_{t=1}^{T} \sum_{m=1}^{M} P(m|\boldsymbol{X}_t, \lambda) P(\boldsymbol{Y}_t|\boldsymbol{X}_t, m, \lambda), \end{aligned} \tag{2.2}$$

where $\boldsymbol{m} = [m_1, m_2, \cdots, m_t, \cdots, m_T]$ indicate mixture component sequence. Maximum likelihood mixture component at frame t $P(m|\boldsymbol{X}_t, \lambda)$ and m-th conditional probability density function $P(\boldsymbol{Y}_t|\boldsymbol{X}_t, m, \lambda)$ follow:

$$P(m \mid \boldsymbol{X}_t, \lambda) = \frac{\alpha_m \mathcal{N}\left(\boldsymbol{X}_t; \boldsymbol{\mu}_m^{(X)}, \boldsymbol{\Sigma}_m^{(XX)}\right)}{\sum_{n=1}^{M} \alpha_n \mathcal{N}\left(\boldsymbol{X}_t; \boldsymbol{\mu}_n^{(X)}, \boldsymbol{\Sigma}_n^{(XX)}\right),} \tag{2.3}$$

$$P(\boldsymbol{Y}_t \mid \boldsymbol{X}_t, m, \lambda) = \mathcal{N}\left(\boldsymbol{Y}_t; \boldsymbol{E}_{m,t}^{(Y)}, \boldsymbol{D}_m^{(Y)}\right), \tag{2.4}$$

where

$$\boldsymbol{E}_{m,t}^{(Y)} = \boldsymbol{\mu}_m^{(Y)} + \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(XX)-1} \left(\boldsymbol{X}_t - \boldsymbol{\mu}_m^{(X)}\right), \tag{2.5}$$

$$\boldsymbol{D}_m^{(Y)} = \boldsymbol{\Sigma}_m^{(YY)} - \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(XX)-1} \boldsymbol{\Sigma}_m^{(XY)}. \tag{2.6}$$

Converted feature vector $\hat{\boldsymbol{y}}$ that maximizes the likelihood function described in Eq. (4.7) is estimated based on following equation.

$$\hat{\boldsymbol{y}} = \mathrm{a}rgmax_{\boldsymbol{y}}\, P(\boldsymbol{Y} \mid \boldsymbol{X}, \lambda) \qquad \text{subject to} \quad \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}, \tag{2.7}$$

where $\boldsymbol{W}$ is a transformation matrix to expand the static feature vector sequence into the joint static and dynamic feature vector sequence [41].

To reduce computational costs of the parameter generation, we approximate the likelihood function of the converted feature vector described in Eq.(4.7) using sub-optimum mixture component sequence $\hat{\boldsymbol{m}} = [\hat{m}_1, \cdots, \hat{m}_T]$ as follows:

$$P(\boldsymbol{Y} \mid \boldsymbol{X}, \lambda) \simeq P(\hat{\boldsymbol{m}} \mid \boldsymbol{X}, \lambda) P(\boldsymbol{Y} \mid \boldsymbol{X}, \hat{\boldsymbol{m}}, \lambda), \tag{2.8}$$

where the sub-optimum mixture component sequence $\hat{\boldsymbol{m}}$ is given by

$$\hat{\boldsymbol{m}} = \underset{\boldsymbol{m}}{\mathrm{a}rgmax}\, P(\boldsymbol{m} \mid \boldsymbol{X}, \lambda). \tag{2.9}$$

Using this sub-optimum mixture component sequence, the converted feature vector is determined as follows:

$$\begin{aligned} \hat{\boldsymbol{y}} &= \underset{\boldsymbol{y}}{\mathrm{a}rgmax}\, P(\hat{\boldsymbol{m}} \mid \boldsymbol{X}, \lambda) P(\boldsymbol{Y} \mid \boldsymbol{X}, \hat{\boldsymbol{m}}, \lambda) \\ &= \left(\boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}} \boldsymbol{E}_{\hat{\boldsymbol{m}}}^{(Y)}, \end{aligned} \tag{2.10}$$

where

$$\boldsymbol{E}_{\hat{\boldsymbol{m}}}^{(Y)} = \left[\boldsymbol{E}_{\hat{m}_1,1}^{(Y)}, \boldsymbol{E}_{\hat{m}_2,2}^{(Y)}, \cdots, \boldsymbol{E}_{\hat{m}_t,t}^{(Y)}, \cdots, \boldsymbol{E}_{\hat{m}_T,T}^{(Y)}\right], \tag{2.11}$$

$$\boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}} = \mathrm{diag}\left[\boldsymbol{D}_{\hat{m}_1}^{(Y)^{-1}}, \boldsymbol{D}_{\hat{m}_2}^{(Y)^{-1}}, \cdots, \boldsymbol{D}_{\hat{m}_t}^{(Y)^{-1}}, \cdots, \boldsymbol{D}_{\hat{m}_T}^{(Y)^{-1}}\right]. \tag{2.12}$$

2.4. Parameter generation considering GV

In order to alleviate the sound quality degradation due to over-smoothing effect of the converted feature trajectory, we incorporate probability density function of global variance (GV) [10] of static feature vector sequence as a constraint term of the likelihood function. The GV of the static feature vector sequence is defined as follows:

$$\boldsymbol{v}(\boldsymbol{y}) = [v(1), v(2), \cdots, v(d), \cdots, v(D)]^\top, \tag{2.13}$$

$$v(d) = \frac{1}{T} \sum_{t=1}^{T} (y_t(d) - \bar{y}(d))^2, \tag{2.14}$$

$$\bar{y}(d) = \frac{1}{T} \sum_{\tau=1}^{T} y_\tau(d), \tag{2.15}$$

where $y_t(d)$ indicates d-th dimensional component of static feature vector of the target speaker at frame t. The probability density function of the converted feature vector considering GV is defined as follows:

$$\begin{aligned} P\left(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\lambda}, \boldsymbol{\lambda}^{(v)}\right) &= P\left(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\lambda}\right)^{e} P\left(\boldsymbol{v}(\boldsymbol{y}) \mid \boldsymbol{\lambda}^{(v)}\right) \\ \text{subject to} \quad \boldsymbol{Y} &= \boldsymbol{W}\boldsymbol{y}, \end{aligned} \tag{2.16}$$

where e is a hyper-parameter that adjusts a balance between two likelihood functions $P\left(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\lambda}\right)$ and $P\left(\boldsymbol{v}(\boldsymbol{y}) \mid \boldsymbol{\lambda}^{(v)}\right)$. Probability density function of the GV is modeled by Gaussian distribution $\boldsymbol{\lambda}^{(v)}$ consisting of mean vector $\boldsymbol{\mu}^{(v)}$ and covariance matrix $\boldsymbol{\Sigma}^{(vv)}$ as follows:

$$P\left(\boldsymbol{v}(\boldsymbol{y}) \mid \boldsymbol{\lambda}^{(v)}\right) = \mathcal{N}\left(\boldsymbol{v}(\boldsymbol{y})\,;\, \boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(vv)}\right). \tag{2.17}$$

The converted feature vector sequence considering GV is estimated by maximizing following objective function,

$$\mathcal{L} = \log\left\{P\left(\boldsymbol{Y} \mid \boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}\right)^{e} P\left(\boldsymbol{v}(\boldsymbol{y}) \mid \boldsymbol{\lambda}^{(v)}\right)\right\}. \tag{2.18}$$

The iterative parameter update using gradient descend is performed as bellow:

$$\begin{aligned} \hat{\boldsymbol{y}}^{(i+1)-th} &= \hat{\boldsymbol{y}}^{(i)-th} + \alpha \cdot \Delta\hat{\boldsymbol{y}}^{(i)-th}, & (2.19) \\ \Delta\boldsymbol{y}^{(i)-th} &= \left.\frac{\partial \mathcal{L}}{\partial \boldsymbol{y}}\right|_{\boldsymbol{y}=\boldsymbol{y}^{(i)-th}}, & (2.20) \end{aligned}$$

where α is a step-size parameter. The first derivative of the objective function is given by

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \boldsymbol{y}} &= e\left(-\boldsymbol{W}^{\top}\boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}}\boldsymbol{W}\boldsymbol{y} + \boldsymbol{W}^{\top}\boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}}\boldsymbol{E}_{\hat{\boldsymbol{m}}}^{(Y)}\right) \\ &\qquad + \left[\boldsymbol{v}_1'^{\top}, \boldsymbol{v}_2'^{\top}, \cdots, \boldsymbol{v}_t'^{\top}, \cdots, \boldsymbol{v}_T'^{\top}\right]^{\top}, & (2.21) \\ \boldsymbol{v}_t' &= \left[v_t'(1), v_t'(2), \cdots, v_t'(d), \cdots, v_t'(D)\right]^{\top}, & (2.22) \\ v_t'(d) &= -\frac{2}{T}\boldsymbol{p}^{(v)}(d)^{\top}\left(\boldsymbol{v}(\boldsymbol{y}) - \boldsymbol{\mu}^{(v)}\right)\left(y_t(d) - \bar{y}(d)\right), & (2.23) \end{aligned}$$

where $\boldsymbol{p}^{(v)}(d)$ indicates a d-th column vector of inverse matrix $\boldsymbol{P}^{(v)} = \boldsymbol{\Sigma}^{(vv)^{-1}}$.

2.5. Post-filtering considering MS

In order to sophisticatedly model the converted feature trajectory compared to GV, modulation spectral (MS) has been proposed [42]. The MS is represented as log-scaled power spectral of the acoustic feature sequence of the target speaker, which is calculated as follows:

$$\begin{aligned} \boldsymbol{s}(\boldsymbol{y}) &= \left[\boldsymbol{s}(1)^\top, \cdots, \boldsymbol{s}(d)^\top, \cdots, \boldsymbol{s}(D)^\top\right]^\top, && (2.24)\\ \boldsymbol{s}(\boldsymbol{d}) &= \left[s_d(0), \cdots, s_d(f), \cdots, s_d(D_s)\right]^\top, && (2.25) \end{aligned}$$

where $s_d(f)$ is the f-th MS of the d-th dimension of the acoustic feature sequence $[y_1(d), \cdots, y_t(d)]^\top$, f is a modulation frequency index, D_s is one half number of the discrete Fourier transform (DFT) length. The MS is calculated from an utterance that is zero-padded to set its length to $2D_s$.

For the training process, two probability density functions of the MS for the natural feature sequence $\boldsymbol{y}$ and converted feature sequence $\hat{\boldsymbol{y}}$ are modeled based on Gaussian distribution $\boldsymbol{\lambda}^{(N)}$ and $\boldsymbol{\lambda}^{(C)}$, respectively. The parameter set $\boldsymbol{\lambda}$ consists of mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}^{(N)}$. The probability density function of MS for the natural feature sequence are given by

$$\begin{aligned} P\left(\boldsymbol{s}(\boldsymbol{y}) \mid \boldsymbol{\lambda}^{(N)}\right) &= \mathcal{N}\left(\boldsymbol{s}(\boldsymbol{y}); \boldsymbol{\mu}^{(N)}, \boldsymbol{\Sigma}^{(N)}\right), && (2.26)\\ \boldsymbol{\mu}^{(N)} &= \left[\mu_{1,0}^{(N)}, \cdots, \mu_{D,D_s}^{(N)}\right]^\top, && (2.27)\\ \boldsymbol{\Sigma}^{(N)} &= \operatorname{diag}\left[\left(\sigma_{1,0}^{(N)}\right)^2, \cdots, \left(\sigma_{D,D_s}^{(N)}\right)^2\right], && (2.28) \end{aligned}$$

where $\mu_{d,f}^{(N)}$ and $\left(\sigma_{d,f}^{(N)}\right)^2$ is a mean and a variance of $s_d(f)$. The probability density function of MS for the converted feature sequence is also given by

$$P\left(\boldsymbol{s}(\hat{\boldsymbol{y}}) \mid \boldsymbol{\lambda}^{(C)}\right) = \mathcal{N}\left(\boldsymbol{s}(\hat{\boldsymbol{y}}); \boldsymbol{\mu}^{(C)}, \boldsymbol{\Sigma}^{(C)}\right). \qquad (2.29)$$

For the filtering process, the following filter is applied to the converted feature sequence $\hat{\boldsymbol{y}}$ as follows:

$$s'_d(f) = (1-k)s_d(f) + k\left[\frac{\sigma_{d,f}^{(N)}}{\sigma_{d,f}^{(C)}}\left(s_d(f) - \mu_{d,f}^{(C)}\right) + \mu_{d,f}^{(N)}\right], \qquad (2.30)$$

where k is a post-filter emphasis parameter valued between 0 through 1.

2.6. Many-to-many VC based on EV-GMM

Many-to-many VC based on EV-GMM is a technique to convert voice timbre of an arbitrary source speaker into that of an arbitrary target speaker [15,43]. Many-to-many VC based on EV-GMM consists of the training process and conversion process.

2.6.1. Training process of EV-GMM

For the training of the EV-GMM, joint feature vectors of a single reference speaker and several pre-stored target speakers are extracted using dynamic time warping in advance in the same manner as VC based on GMM. Time sequence vectors of the reference speaker and the pre-stored target speaker are denoted as $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$, $\boldsymbol{Y}_t^{(s)} = [\boldsymbol{y}_t^{(s)\top}, \Delta\boldsymbol{y}_t^{(s)\top}]^\top$ where T is the number of frames included in the time sequence of the given source feature vectors. The number of pre-stored target speakers is S. Joint probability density function of the EV-GMM is given by

$$P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)} | \boldsymbol{\lambda}^{(EV)}, \boldsymbol{e}^{(s)}\right) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix} \boldsymbol{X}_t \\ \boldsymbol{Y}_t^{(s)} \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)}(s) \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_m^{(YY)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix}\right), \tag{2.31}$$

where the mean vector of s-th pre-stored target speaker follows:

$$\boldsymbol{\mu}_m^{(Y)}(s) = \boldsymbol{A}_m \boldsymbol{e}^{(s)} + \boldsymbol{l}_m, \tag{2.32}$$

where $\boldsymbol{e}^{(s)} = [e^{(s)}(1), \cdots, e^{(s)}(J)]^\top$ is a J-dimensional eigen weight vector of s-th pre-stored target speaker. $\boldsymbol{\lambda}^{(EV)}$ is a EV-GMM parameter set consisting of basis vector $\boldsymbol{A}_m = [\boldsymbol{a}_{m,1}, \cdots, \boldsymbol{a}_{m,J}]$ and bias vector $\boldsymbol{l}_m = [l_{m,1}, \cdots, l_{m,J}]$ depending on the mixture component in addition to the parameter set of the GMM.

The training process of the EV-GMM is composed of training processes of speaker-independent GMM (SI-GMM) and speaker-dependent GMM (SD-GMM), principal component analysis and speaker adaptive training (SAT). Figure 2.5 indicates the training process of the EV-GMM.

***Step*1 : Training process of SI-GMM**

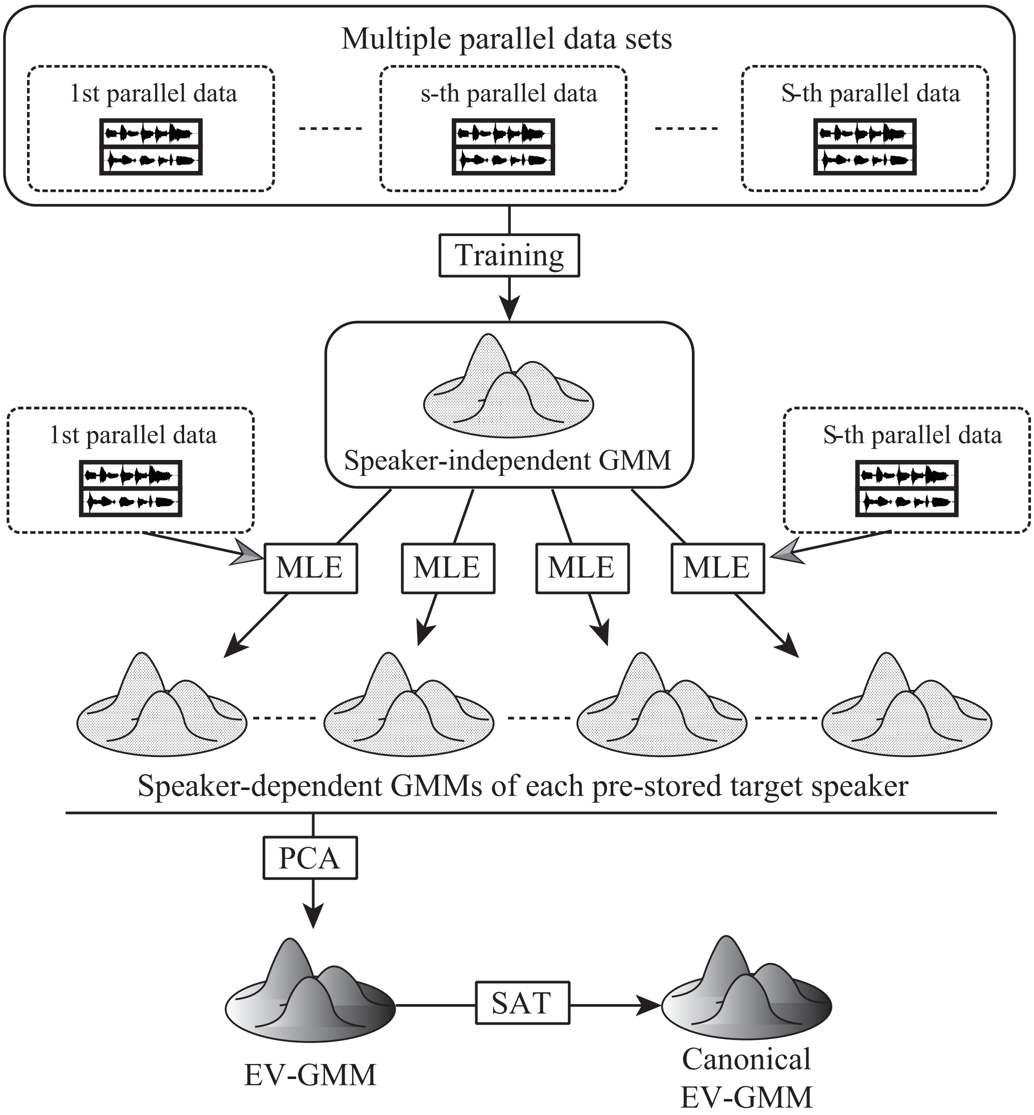


Figure 2.5.: Training process of EV-GMM.

Joint probability density function of the reference speaker and all pre-stored target speakers is given by

$$\hat{\boldsymbol{\lambda}}^{(0)} = \underset{\boldsymbol{\lambda}}{argmax} \prod_{s=1}^{S} \prod_{t=1}^{T_s} P(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)} | \boldsymbol{\lambda}), \tag{2.33}$$

where T_s is the frame number of joint feature vector for the reference speaker and s-th pre-stored target speaker. $\boldsymbol{\lambda}^{(0)}$ indicates a parameter set of the SI-GMM consisting of the mixture-component weight α_m, the mean vector $\boldsymbol{\mu}_m$, and the covariance matrix $\boldsymbol{\Sigma}_m$ of the m-th mixture component.

***Step*2 : Training process of SD-GMM**

For the training process of the SD-GMM of s-the pre-stored target speaker, the mean vector of the SI-GMM $\boldsymbol{\lambda}^{(0)}$ is updated as follow:

$$\hat{\boldsymbol{\lambda}}^{(s)} = \underset{\boldsymbol{\lambda}}{\mathrm{a}rgmax} \prod_{t=1}^{T_s} P(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)} | \boldsymbol{\lambda}^{(0)}), \tag{2.34}$$

where $\hat{\boldsymbol{\lambda}}^{(s)}$ is a parameter set of the SD-GMM of s-th pre-stored target speaker.

Let $\boldsymbol{Y}_t(s) = [\boldsymbol{Y}_t^\top(s), \Delta\boldsymbol{Y}_t^\top(s)]^\top$ denote the joint static and delta feature vector at frame t of the pre-stored target speaker s to be adapted. The mean vector set of the s-th pre-stored target speaker $\hat{\boldsymbol{\mu}}(s) = \{\hat{\boldsymbol{\mu}}_1(s), \cdots, \hat{\boldsymbol{\mu}}_M(s)\}$ is given by maximizing following likelihood function:

$$\hat{\boldsymbol{\mu}}(s) = \underset{\boldsymbol{\mu}(s)}{\mathrm{a}rgmax} \prod_{t=1}^{T} P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t(s) | \lambda^{(0)}, \boldsymbol{\mu}(s)\right). \tag{2.35}$$

This update process is performed using the expectation-maximization (EM) algorithm using following auxiliary function:

$$\begin{aligned} Q\left(\boldsymbol{\mu}(s), \hat{\boldsymbol{\mu}}(s)\right) &= \sum_{t=1}^{T} \sum_{m=1}^{M} P\left(m | \boldsymbol{X}_t, \boldsymbol{Y}_t(s), \lambda^{(0)}, \boldsymbol{\mu}_m(s)\right) \\ &\quad \log P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t(s), m | \lambda^{(0)}, \hat{\boldsymbol{\mu}}_m(s)\right). \end{aligned} \tag{2.36}$$

The maximum likelihood estimate of the m-th target mean vector $\hat{\boldsymbol{\mu}}_m(s)$ is calculated as follows:

$$\begin{aligned} \hat{\boldsymbol{\mu}}_m(s) = &\left\{ \sum_{m=1}^{M} \Gamma_m \boldsymbol{P}_m^{(YY)} \right\}^{-1} \\ &\left\{ \sum_{m'=1}^{M} \boldsymbol{P}_{m'}^{(YY)} \overline{\boldsymbol{Y}}_{m'}(s) + \boldsymbol{P}_{m'}^{(YX)} \left(\overline{\boldsymbol{X}}_{m'} - \Gamma_{m'} \boldsymbol{\mu}_{m'}^{(X)} \right) \right\}, \end{aligned} \tag{2.37}$$

where

$$
\begin{aligned}
\Gamma_m &= \sum_{t=1}^{T} P\left(m|\boldsymbol{X}_t, \boldsymbol{Y}_t(s), \lambda^{(0)}, \boldsymbol{\mu}_m(s)\right), && (2.38)\\
\overline{\boldsymbol{Y}}_m(s) &= \sum_{t=1}^{T} P\left(m|\boldsymbol{X}_t, \boldsymbol{Y}_t(s), \lambda^{(0)}, \boldsymbol{\mu}_m(s)\right) \boldsymbol{Y}_t(s), && (2.39)\\
\overline{\boldsymbol{X}}_m &= \sum_{t=1}^{T} P\left(m|\boldsymbol{X}_t, \boldsymbol{Y}_t(s), \lambda^{(0)}, \boldsymbol{\mu}_m(s)\right) \boldsymbol{X}_t, && (2.40)
\end{aligned}
$$

$$
\begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{P}_m^{(XX)} & \boldsymbol{P}_m^{(XY)} \\ \boldsymbol{P}_m^{(YX)} & \boldsymbol{P}_m^{(YY)} \end{bmatrix}. \qquad (2.41)
$$

$Step3$: Principal component analysis

$2DM$-dimensional super vector $\boldsymbol{SV}^{(s)} = \left[\boldsymbol{\mu}_1^{(Y)}(S)^\top, \cdots, \boldsymbol{\mu}_M^{(Y)}(S)^\top\right]^\top$ is obtained by concatenating M mean vectors in each pre-stored target speaker. The basis vector $\boldsymbol{A}$ and bias vector $\boldsymbol{l}$ are derived based on principal component analysis using the super vector as bellow

$$
\begin{aligned}
\boldsymbol{SV}^{(s)} &\approx [\boldsymbol{A}_1^\top, \cdots, \boldsymbol{A}_M^\top]^\top \boldsymbol{e}^{(s)} + [\boldsymbol{l}_1^\top, \cdots, \boldsymbol{l}_M^\top], && (2.42)\\
\boldsymbol{l}_m &= \frac{1}{S} \sum_{s=1}^{S} \boldsymbol{\mu}_m^{(Y)}(s). && (2.43)
\end{aligned}
$$

$Step4$: Speaker adaptive training

To improve the modeling accuracy of the EV-GMM, further optimization of the EV-GMM using speaker adaptive training (SAT) [44,45] is performed [43]. The EV-GMM is refined by maximizing following likelihood function using a parallel data set in each pre-stored target speaker as follows:

$$
\left\{\hat{\lambda}^{(EV)}, \hat{\boldsymbol{\Omega}}^{(S)}\right\} = \underset{\boldsymbol{\lambda}^{(EV)}, \boldsymbol{\Omega}^{(S)}}{\mathrm{a}rgmax} \prod_{s=1}^{S} \prod_{t=1}^{T_s} P(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)} | \boldsymbol{\lambda}^{(EV)}, \boldsymbol{e}^{(s)}), \qquad (2.44)
$$

where $\boldsymbol{\Omega}^{(S)} = \left\{\boldsymbol{e}^{(1)}, \ldots, \boldsymbol{e}^{(s)}, \ldots, \boldsymbol{e}^{(S)}\right\}$ is a set of eigen weight vectors of all pre-stored target speakers. In the optimization process based on SAT, following objective function is maximized based on EM algorithm.

$$\begin{aligned} Q\left(\left\{\hat{\boldsymbol{\lambda}}^{(EV)}, \hat{\boldsymbol{\Omega}}^{(S)}\right\}, \left\{\boldsymbol{\lambda}^{(s)}, \boldsymbol{\Omega}^{(S)}\right\}\right) &= \sum_{s=1}^{S}\sum_{t=1}^{T_s}\sum_{m=1}^{M} P(m|\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}, \boldsymbol{\lambda}^{(EV)}, \boldsymbol{e}^{(s)}) \\ &\quad \log P(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}, m|\boldsymbol{\lambda}^{(EV)}, \boldsymbol{e}^{(s)}). \end{aligned} \tag{2.45}$$

In the E-step, posterior probability $P(m|\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}, \boldsymbol{\lambda}^{(EV)}, \boldsymbol{e}^{(s)})$ is estimated. Then, in the M-step, the EV-GMM parameter set depends on each pre-stored target speakers is updated. In the M-step, the parameters are updated one-by-one because it is difficult to update simultaneously.

At fast, eigen weight vector in each pre-stored target speaker is updated. Maximum likelihood estimate of the eigen weight vector of s-th pre-stored target speaker is given by

$$\begin{aligned} \hat{\boldsymbol{e}}^{(s)} &= \left(\sum_{m=1}^{M} \Gamma_{m,s} \boldsymbol{A}_m^\top \boldsymbol{P}_m^{(YY)} \boldsymbol{A}_m\right)^{-1} \\ &\quad \left[\sum_{m=1}^{M} \boldsymbol{A}_m^\top \left\{\boldsymbol{P}_m^{(YX)}(\overline{\boldsymbol{X}}_m^{(s)} - \Gamma_m^{(s)}\boldsymbol{\mu}_m^{(X)}) + \boldsymbol{P}_m^{(YY)}(\overline{\boldsymbol{Y}}_m^{(s)} - \Gamma_m^{(s)}\boldsymbol{l}_m)\right\}\right], \end{aligned} \tag{2.46}$$

where

$$\Gamma_m^{(s)} = \sum_{t=1}^{T_S} \gamma_{m,t}^{(s)} = \sum_{t=1}^{T_S} P(m|\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}, \boldsymbol{\lambda}^{(EV)}, \boldsymbol{e}^{(s)}), \tag{2.47}$$

$$\overline{\boldsymbol{X}}_m^{(s)} = \sum_{t=1}^{T_S} \gamma_{m,t}^{(s)} \boldsymbol{X}_t, \tag{2.48}$$

$$\overline{\boldsymbol{Y}}_m^{(s)} = \sum_{t=1}^{T_S} \gamma_{m,t}^{(s)} \boldsymbol{Y}_t^{(s)}, \tag{2.49}$$

$$\boldsymbol{\Sigma}_m^{(X,Y)^{-1}} = \begin{bmatrix} \boldsymbol{P}_m^{(XX)} & \boldsymbol{P}_m^{(XY)} \\ \boldsymbol{P}_m^{(YX)} & \boldsymbol{P}_m^{(YY)} \end{bmatrix}. \tag{2.50}$$

Next, mean vector of the EV-GMM is updated as follows:

$$\hat{\boldsymbol{v}}_m = \left(\sum_{s=1}^{S} \Gamma_m^{(s)} \hat{\boldsymbol{W}}_s^\top \boldsymbol{\Sigma}_m^{(XY)^{-1}} \hat{\boldsymbol{W}}_s\right)\left(\sum_{s=1}^{S} \Gamma_m^{(s)} \hat{\boldsymbol{W}}_s^\top \boldsymbol{\Sigma}_m^{(XY)^{-1}} \overline{\mathbf{Z}}_m^{(s)}\right), \tag{2.51}$$

where

$$\overline{\boldsymbol{Z}}_m^{(s)} = \left[\overline{\boldsymbol{X}}_m^{(s)\top}, \overline{\boldsymbol{Y}}_m^{(s)\top}\right]^\top, \tag{2.52}$$

$$\hat{\boldsymbol{v}}_m = \left[\hat{\boldsymbol{\mu}}_m^{(X)\top}, \hat{\boldsymbol{l}}_m^\top, \hat{\boldsymbol{a}}_m(1)^\top, \hat{\boldsymbol{a}}_m(2)^\top, \cdots, \hat{\boldsymbol{a}}_m(J)^\top\right]^\top, \tag{2.53}$$

$$\hat{\boldsymbol{W}}_s = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} & \hat{e}_1^{(s)}\boldsymbol{I} & \hat{e}_2^{(s)}\boldsymbol{I} & \cdots & \hat{e}_J^{(s)}\boldsymbol{I} \end{bmatrix}. \tag{2.54}$$

Then, mixture component weights of the EV-GMM are updated as follows:

$$\hat{\alpha}_m = \frac{\sum_{s=1}^{S} \Gamma_m^{(s)}}{\sum_{m=1}^{M}\sum_{s=1}^{S} \Gamma_m^{(s)}}. \tag{2.55}$$

Finally, covariance matrix of EV-GMM is updated as bellow:

$$\boldsymbol{\Sigma}_m^{(X,Y)} = \frac{1}{\sum_{s=1}^{S} \Gamma_m^{(s)}} \sum_{s=1}^{S} \left\{ \overline{\boldsymbol{V}}_m^{(s)} + \Gamma_m^{(s)} \hat{\boldsymbol{\mu}}_{m,s}^{(X,Y)} \hat{\boldsymbol{\mu}}_{m,s}^{(X,Y)\top} - \left(\hat{\boldsymbol{\mu}}_{m,s}^{(X,Y)} \overline{\boldsymbol{Z}}_m^{(s)\top} + \overline{\boldsymbol{Z}}_m^{(s)} \hat{\boldsymbol{\mu}}_{m,s}^{(X,Y)\top} \right) \right\}, \tag{2.56}$$

where

$$\overline{\boldsymbol{V}}_m^{(s)} = \sum_{t=1}^{T_s} \gamma_{m,t}^{(s)} \left[\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^{(s)\top}\right] \left[\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^{(s)\top}\right]^\top, \tag{2.57}$$

$$\hat{\boldsymbol{\mu}}_{m,s}^{(X,Y)} = \hat{\boldsymbol{W}}_s \overline{\boldsymbol{V}}_m^{(s)} = \begin{bmatrix} \hat{\boldsymbol{\mu}}_m^{(X)} \\ \hat{\boldsymbol{A}}_m \hat{\boldsymbol{e}}^{(s)} + \hat{\boldsymbol{l}}_m \end{bmatrix}. \tag{2.58}$$

2.6.2. Conversion process

Figure 2.6 indicates adaptation process of arbitrary source and target speakers and conversion process based on the EV-GMM. In the conversion process, the joint probability density function of the acoustic feature between the source and target speakers is derived as

$$
\begin{aligned}
&P\left(\boldsymbol{Y}_t^{(i)}, \boldsymbol{Y}_t^{(o)} | \boldsymbol{\lambda}^{(EV)}, e^{(i)}, e^{(o)}\right) \\
&= \sum_{m=1}^{M} P\left(m|\boldsymbol{\lambda}^{(EV)}\right) \int P\left(\boldsymbol{Y}_t^{(i)} | \boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(EV)}, e^{(i)}\right) \\
&\quad P\left(\boldsymbol{Y}_t^{(o)} | \boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(EV)}, e^{(o)}\right) P\left(\boldsymbol{X}_t | m, \boldsymbol{\lambda}^{(EV)}\right) \mathrm{d}\boldsymbol{X}_t \\
&= \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix} \boldsymbol{Y}_t^{(i)} \\ \boldsymbol{Y}_t^{(o)} \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu}_m^{(Y)}(i) \\ \boldsymbol{\mu}_m^{(Y)}(o) \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_m^{(YY)} & \boldsymbol{\Sigma}_m^{(YXY)} \\ \boldsymbol{\Sigma}_m^{(YXY)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix}\right), \qquad (2.59)
\end{aligned}
$$

$$
\boldsymbol{\Sigma}_m^{(YXY)} = \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(XX)^{-1}} \boldsymbol{\Sigma}_m^{(XY)}, \qquad (2.60)
$$

where $\boldsymbol{Y}_t^{(i)}$ and $\boldsymbol{Y}_t^{(o)}$ indicate the source and target speakers' static and dynamic feature vectors, respectively. Speaker-dependent models of the source and target speaker $\boldsymbol{\mu}_m^{(Y)}(i)$ and $\boldsymbol{\mu}_m^{(Y)}(o)$ are determined by Eq. (2.61) by the use of source and target speakers' eigen weight parameters $e^{(i)}$, $e^{(o)}$, respectively. Their eigen weight parameters $e^{(i)}$, $e^{(o)}$ are given by

$$
\begin{aligned}
\hat{\boldsymbol{e}} &= \underset{\boldsymbol{e}}{\mathrm{a}rgmax} \int P(\boldsymbol{X}, \boldsymbol{Y}^{(tar)} | \boldsymbol{\lambda}^{(EV)}, \boldsymbol{e}) d\boldsymbol{X} \\
&= \underset{\boldsymbol{e}}{\mathrm{a}rgmax} \prod_{t=1}^{T} \int P(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(tar)} | \boldsymbol{\lambda}^{(EV)}, \boldsymbol{e}) d\boldsymbol{X}_t \\
&= \underset{\boldsymbol{e}}{\mathrm{a}rgmax} \prod_{t=1}^{T} P(\boldsymbol{Y}_t^{(tar)} | \boldsymbol{\lambda}^{(EV)}, \boldsymbol{e}), \qquad (2.61)
\end{aligned}
$$

where $\boldsymbol{Y}^{(tar)}$ indicates feature vector sequence of the arbitrary speaker. Using this many-to-many EV-GMM, the converted voice is estimated based on maximum likelihood parameter generation in the same manner as described in Section 2.3.2.

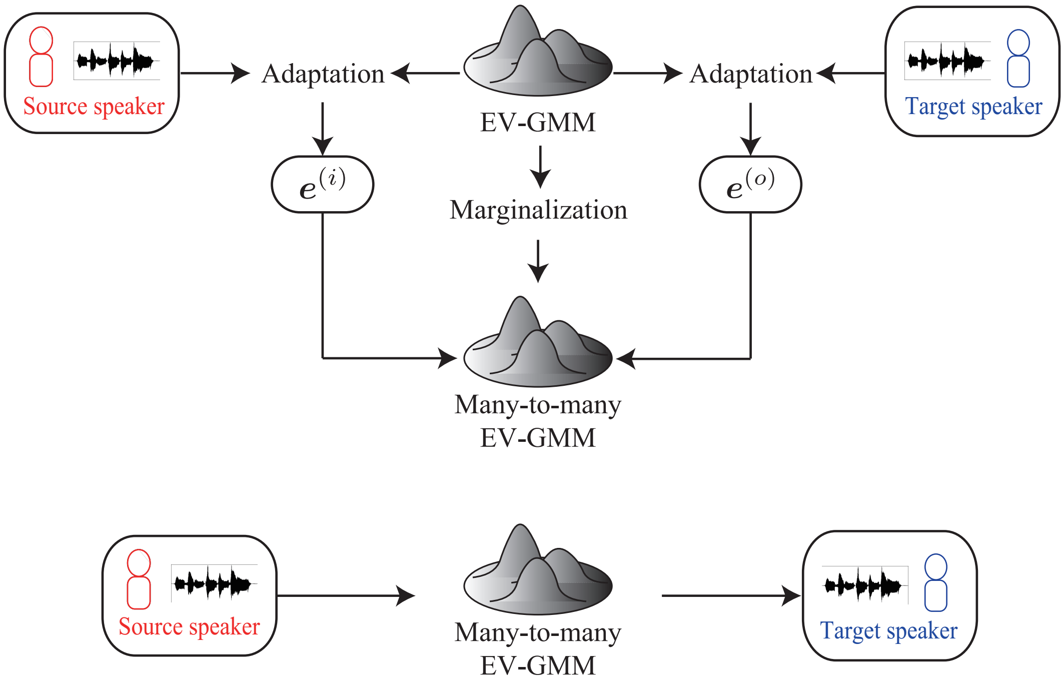


Figure 2.6.: Conversion process of many-to-many VC based on EV-GMM.

2.7. Voice timbre control based on MR-GMM

2.7.1. Training process of MR-GMM

Voice timbre control based on the MR-GMM is a technique to control voice timbre of the source speaker using perceptually understandable voice timbre expression words. VC based on the MR-GMM consists of the training process and conversion process. The MR-GMM is also trained using multiple parallel data sets consisting of a single reference's singing voices and many pre-stored target speakers' voices. The joint probability density of $2D$-dimensional joint static and dynamic feature vectors modeled by the MR-GMM is given by

$$P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}|\boldsymbol{\lambda}^{(MR)}, \boldsymbol{w}^{(s)}\right) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix} \boldsymbol{X}_t \\ \boldsymbol{Y}_t^{(s)} \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)}(s) \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix}\right), \tag{2.62}$$

where $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$ and $\boldsymbol{Y}_t^{(s)} = [\boldsymbol{Y}_t^{(s)\top}, \Delta\boldsymbol{Y}_t^{(s)\top}]^\top$ show static and delta feature vectors of the source and s-th pre-stored target speaker. The mean vector of the s-th pre-stored

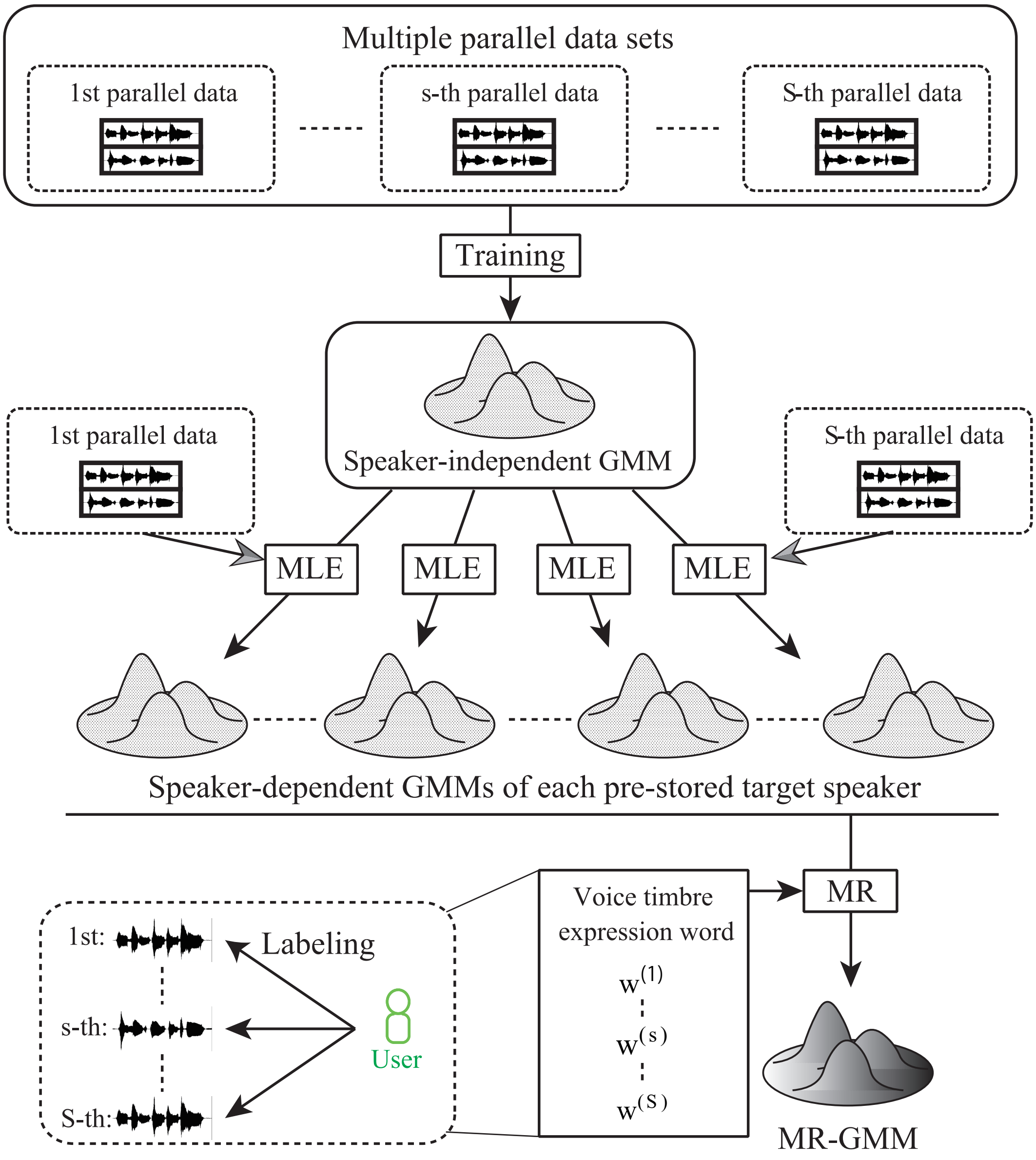


Figure 2.7.: Training process of the MR-GMM.

speaker is given by

$$\boldsymbol{\mu}_m^{(Y)}(s) = \boldsymbol{b}_m^{(Y)} \boldsymbol{w}^{(s)} + \overline{\boldsymbol{\mu}}_m^{(Y)}, \tag{2.63}$$

where $\boldsymbol{b}_m^{(Y)}$ and $\overline{\boldsymbol{\mu}}_m^{(Y)}$ indicate the representative vector and bias vector respectively. $\boldsymbol{w}^{(s)}$ indicates the s-th pre-stored target speaker's perceived score, which is manually assigned in each pre-stored target speaker in advance.

In this training process, SD-GMM modeling is performed in the same manner as the training process of EV-GMM described in Section. 2.6.1. Then, to model the representative and bias vectors, the multiple-regression analysis is performed.

2.7.2. Conversion process

In the conversion process, voice timbre expression scores $\boldsymbol{w}$ are manually set to the desired value to determine the target mean vector. Then, the converted feature vector is estimated in the same manner as described in Section 2.3.2.

2.8. Low-delay conversion algorithm in VC based on GMM

2.8.1. Conversion process

In the maximum likelihood parameter generation described in Eq. 2.10, the converted feature vector is estimated considering 1-st order derivation components based on the static and dynamic transformation matrix $\boldsymbol{W}$. Although this parameter generation technique makes it possible to estimate the converted feature vector smoothly, it is difficult to directly apply it into a real-time conversion system because the converted feature vector is estimated base on batch type conversion algorithm. In order to estimate the converted feature vector based on the frame-by-frame manner, a low-delay conversion algorithm has been proposed [24, 25].

Equation 2.10 can be written as follows:

$$\begin{aligned} \boldsymbol{R}\hat{\boldsymbol{y}} &= \boldsymbol{r}, & (2.64)\\ \boldsymbol{R} &= \boldsymbol{P}^{-1} = \boldsymbol{W}^{\top}\boldsymbol{D}_{\hat{m}}^{(Y)^{-1}}\boldsymbol{W}, & (2.65)\\ \boldsymbol{r} &= \boldsymbol{W}^{\top}\boldsymbol{D}_{\hat{m}}^{(Y)^{-1}}\boldsymbol{E}_{m}^{(Y)}. & (2.66) \end{aligned}$$

Using only diagonal components of the covariance matrix $\boldsymbol{\Sigma}_m^{(Y|X)}$, each dimensional components of $\hat{\boldsymbol{y}}$ is separately determined. A $(L+1)$-by-$(L+1)$ state covariance matrix

$\boldsymbol{P}_d^{(0)}$ and a $(L+1)$-dimensional state vector $\hat{\boldsymbol{y}}_d^{(0)}$ are initialized as the zero matrix and the zero vector, respectively. Then, they are recursively updated frame by frame as follows:

$$\begin{aligned}
\boldsymbol{P}_d^{\prime(t-1)} &= \boldsymbol{J}_L \boldsymbol{P}_d^{(t-1)} \boldsymbol{J}_L^\top + \mathrm{diag}\left[\boldsymbol{0}_{1\times L}, \Sigma_{m_t,d}^{(y|X)}\right], && (2.67)\\
\boldsymbol{y}_d^{\prime(t-1)} &= \boldsymbol{J}_L \boldsymbol{y}_d^{(t-1)} + \mathrm{diag}\left[\boldsymbol{0}_{1\times L}, \mu_{m_t,t,d}^{(y|X)}\right]^\top, && (2.68)\\
\boldsymbol{P}_d^{(t)} &= \left(\boldsymbol{I} - \boldsymbol{k}_d^{(t)} \boldsymbol{w}_L\right) \boldsymbol{P}_d^{\prime(t-1)}, && (2.69)\\
\hat{\boldsymbol{y}}_d^{(t)} &= \hat{\boldsymbol{y}}_d^{\prime(t-1)} + \boldsymbol{k}_d^{(t)} \left(\mu_{m_t,t,d}^{(\Delta y|X)} - \boldsymbol{w}_L \hat{\boldsymbol{y}}_d^{\prime(t-1)}\right), && (2.70)\\
& && (2.71)
\end{aligned}$$

where the $(L+1)$-dimensional vector $\boldsymbol{k}_d^{(t)}$ is calculated as

$$\boldsymbol{k}_d^{(t)} = \boldsymbol{P}_d^{(t-1)} \boldsymbol{w}_t^\top \left(\Sigma_{m_t,d}^{(\Delta y|X)} + \boldsymbol{w}_L \boldsymbol{P}_d^{(t-1)} \boldsymbol{w}_L^\top\right)^{-1}, \tag{2.72}$$

where the $(L+1)$-dimensional row vector $\boldsymbol{w}_L$ and the $(L+1)$-by-$(L+1)$ matrix $\boldsymbol{J}_L$ are given by

$$\begin{aligned}
\boldsymbol{w}_L &= \left[\boldsymbol{0}_{1\times(L-1)}, -1, 1\right], && (2.73)\\
\boldsymbol{J}_L &= \begin{bmatrix} 0 & \boldsymbol{I}_{L\times L} \\ 0 & \boldsymbol{0}_{1\times L} \end{bmatrix}, && (2.74)
\end{aligned}$$

respectively. The d-th dimensional static feature components, $\mu_{m,t,d}^{(y|X)}$ and $\Sigma_{m,d}^{(y|X)}$, of the mean vector $\boldsymbol{\mu}_{m,t}^{(Y|X)}$ and the covariance matrix $\boldsymbol{\Sigma}_m^{(Y|X)}$ are used to predict the state covariance matrix and the state vector as shown in Eqs 2.67 and 2.68. Their dynamic feature components, $\mu m_t, t, d^{(\Delta y|X)}$ and $\Sigma_{m_t,d}^{(\Delta y|X)}$, are used to optimize the Kalman gain in Eq. 2.72 and update the state covariance matrix and the state vector as shown in Eqs. 2.69 and 2.70. The first components of $\hat{\boldsymbol{y}}_d^{(t)}$ is used as the d-th component of the converted static feature vector at frame $t-L$, $\hat{y}_{t-L,d}$.

2.8.2. Frame-based GV post-filter for low delay conversion

In order to alleviate sound quality degradation caused by the over-smoothing effect in the low-delay conversion, a frame-based GV post-filter technique has been proposed [25]. The converted static feature vector is enhanced based on frame-based GV as follows:

$$\hat{y}^{(GV)}_{t,d} = \mu^{(v)\frac{1}{2}}_{d} \hat{\mu}^{(v)-\frac{1}{2}}_{d} \left(\hat{y}_{t,d} - \bar{y}_{d}\right) + \bar{y}_{d}, \tag{2.75}$$

where $\bar{\mu}^{(v)}_{d}$ and $\bar{y}_{d}$ denote GV and mean vector of d-th dimensional converted static feature vector without considering GV, which are previously calculated.

2.9. Issues of conventional VC frameworks

2.9.1. Sound quality degradation caused by vocoding

In the VC based on GMM, the vocoding process is necessary to perform to generate a waveform signal of the converted voice. The generate waveform using the vocoding usually causes sound quality degradation even when using original acoustic features extracted from the original waveform because vocoding is a lossy process. Therefore, the sound quality degradation is unavoidable in the conventional statistical VC frameworks. To make it possible to improve the sound quality of the original waveform, it is ineluctable to improve the vocoding framework or avoid the use of the vocoding framework.

2.9.2. Difficulty to retain speaker/singer individuality in voice timbre control based on the MR-GMM

In addition to the problem of the vocoding framework described in 2.9.1, in the voice timbre control based on MR-GMM, it is difficult to retain a speaker/singer individuality after conversion. This problem mainly comes from a representation of the target mean vector of the MR-GMM. The target mean vector is represented by the averaged voice timbre of the target voice timbre scores. Therefore, regardless of the source speaker, the source voice is converted into an identical voice timbre unless changing the target voice timbre score.

2.10. Summary

In this chapter, we describe conventional VC frameworks.

Section 2.2: This section illustrates analysis/synthesis process of a speech waveform based on vocoding.

Section 2.3: This section describes training process and conversion process of VC based on GMM.

Section 2.6: In this section, we describe many-to-many VC based on EV-GMM and adaptation techniques for an arbitrary source/target speaker.

Section 2.7: In this section, we described voice timbre control technique based on MR-GMM.

Section 2.8: This section describes low-delay conversion algorithm for GMM-based VC frameworks.

Section 2.9: In this section, we describe issues of the conventional VC techniques.

3. Voice conversion via statistical waveform modification

3.1. Introduction

This chapter presents a statistical voice conversion (VC) technique for both speaking and singing voice with direct waveform modification based on the spectrum differential (DIFFVC) that can convert voice timbre of a source speaker/singer into that of a target speaker/singer without waveform generation of the converted voice based on vocoding. VC makes it possible to convert voice characteristics of an arbitrary source speaker/singer into those of an arbitrary target speaker/singer by converting several acoustic features such as F_0, aperiodicity, and spectral features based on statistical conversion function. However, the sound quality of the converted voice is usually degraded compared with that of a natural voice due to various factors such as analysis and modeling errors in the vocoding process and over-smoothing effect of converted feature trajectory. To alleviate the sound quality degradation, in this chapter, we propose a statistical waveform modification technique that directly modifies the signal in the waveform domain by estimating the difference in the spectra of the source and target speaker/singers' voices.

In singing VC (SVC) based on the Gaussian mixture model (GMM), F_0, aperiodicity, and spectral envelopes are extracted from the source singer's singing voice and converted to those of the target singer. On the other hand, regarding intra-gender SVC such as male-to-male and female-to-female singer individuality conversions, it is not always necessary to transform F_0 values of the source singer to those of the target because both singers often sing on the same key. Moreover, the conversion of the aperiodicity usually causes only a small impact on the converted singing voice within intra-gender singer pairs. Therefore, it is expected that only spectral conversion is

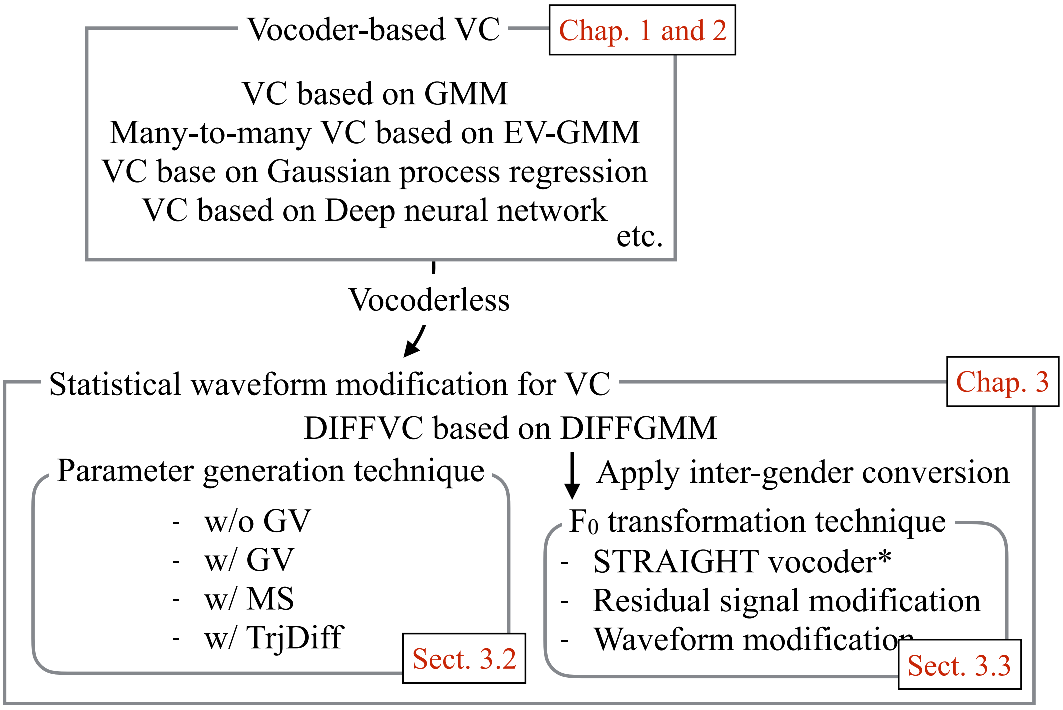

Vocoder-based VC
Chap. 1 and 2
VC based on GMM
Many-to-many VC based on EV-GMM
VC base on Gaussian process regression
VC based on Deep neural network
etc.
Vocoderless
Statistical waveform modification for VC
Chap. 3
DIFFVC based on DIFFGMM
Parameter generation technique
- w/o GV
- w/ GV
- w/ MS
- w/ TrjDiff
Sect. 3.2
Apply inter-gender conversion
F0 transformation technique
- STRAIGHT vocoder*
- Residual signal modification
- Waveform modification
Sect. 3.3

Figure 3.1.: The rest of Chapter 3. *Although STRAIGHT vocoder actually uses vocoding framework, it is one kind of statistical waveform modification.

sufficient to achieve acceptable quality in intra-gender SVC. Based on this idea, in the proposed SVC method, we only focus on converting the spectral envelopes. The waveform of the source singer is directly modified with a digital filter that uses the time-varying difference in the spectral envelope between the source and target singer's singing voices. Note that this spectrum differential is statistically estimated from the spectral envelopes of the source singer. For the intra-gender SVC based on direct waveform modification using spectral differential (DIFFSVC), we propose following techniques: 1) derivation of a differential GMM (DIFFGMM), 2) parameter generation algorithm considering global variance (GV), 2) parameter generation algorithm considering modulation spectral (MS), and 3) parameter generation algorithm based on trajectory differential feature.

For inter-gender conversion such as male-to-female and female-to-male conversions, we propose several F_0 transformation techniques for VC with direct waveform modification with spectral differential (DIFFVC) for normal speech to make it possible to widely accept various F_0 transformation ratios. It is not straightforward to apply the intra-gender DIFFSVC method to intra/inter-gender VC because more complicated F_0 transformation is necessary for VC compared with SVC; e.g., even if using a simple F_0 transformation method with a constant F_0 transformation ratio [46], such a ratio widely varies depending on a combination of the source and target speakers although it can be fixed to double or half in inter-gender SVC which is corresponding to a single key. The following F_0 transformation techniques with or without using vocoding are proposed: 1) DIFFVC with F_0 transformation using STRAIGHT vocoder, 2) DIFFVC with F_0 transformation based on the residual signal modification using time-scaling and resampling, and 3) DIFFVC with F_0 transformation based on waveform modification using time-scaling and resampling.

This chapter is organized as shown in Figure 3.1. Intra-gender DIFFSVC framework including several parameter generation techniques is described in Section 3.2. Intra/inter-gender DIFFVC framework with several F_0 transformation techniques is described in Section 3.3. The experimental evaluations of the proposed methods are described in Section 3.4. The experimental results are briefly summarized in Section 3.5. This chapter is summarized in Section 3.6.

3.2. Intra-gender DIFFSVC based on DIFFGMM

Figures 3.2 (a) and (b) show the conversion processes of the proposed DIFFSVC methods. In the conventional conversion process described in Figure 2.4, the sound quality of the converted singing voice is usually degraded compared with that of the natural singing voice due to F_0 extraction errors, unvoiced/voiced decision errors, spectral parameterization errors caused by liftering, which are brought from analysis and synthesis process of source singing voice. These errors are difficult to avoid even if using high-quality vocoding frameworks. To avoid the sound quality degradation of the converted voice caused by the vocoding in SVC, we propose a statistical waveform modification technique using time-variant spectral feature differential that can avoid using vocoding. In the proposed conversion process, the difference of the spectral features between the source and target singers is estimated from the source singer's spectral features using either a differential GMM (DIFFGMM) or intra/inter-singer GMMs. Voice timbre of the source singer is converted into that of the target singer by directly

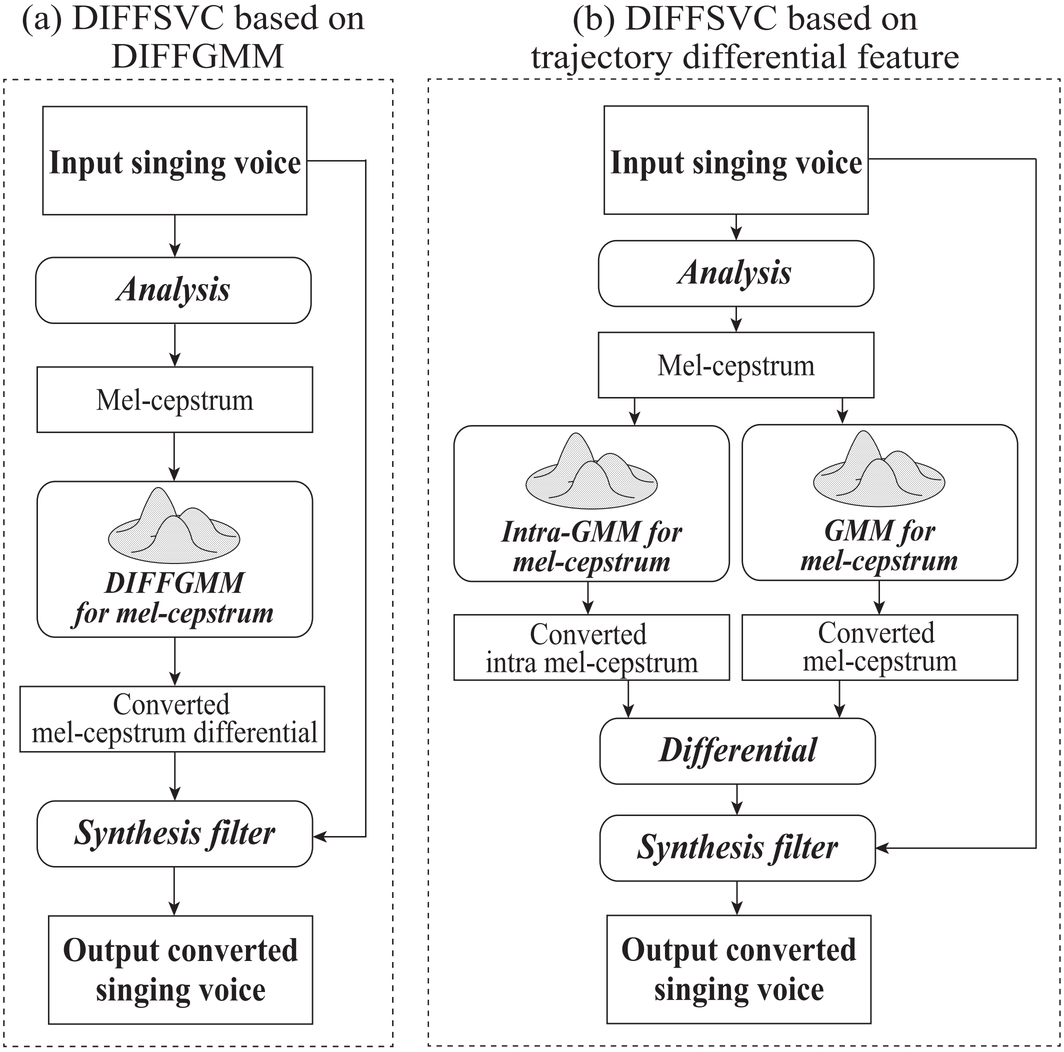

(a) DIFFSVC based on DIFFGMM
Input singing voice
Analysis
Mel-cepstrum
DIFFGMM for mel-cepstrum
Converted mel-cepstrum differential
Synthesis filter
Output converted singing voice
(b) DIFFSVC based on trajectory differential feature
Input singing voice
Analysis
Mel-cepstrum
Intra-GMM for mel-cepstrum
GMM for mel-cepstrum
Converted intra mel-cepstrum
Converted mel-cepstrum
Differential
Synthesis filter
Output converted singing voice

Figure 3.2.: Conversion processes of DIFFSVC based on DIFFGMM and DIFFVC based on trajectory differential.

filtering an input natural singing voice of the source singer with the converted spectral feature differential. The proposed conversion process does not need any waveform generation using vocoding because an original waveform of the source singing voice is directly used as an excitation signal. Therefore, the converted singing voice is free from various errors usually observed in the conventional SVC using waveform generation based on vocoding, such as F_0 extraction errors, unvoiced/voiced decision errors, spectral parameterization errors caused by liftering on the mel-cepstrum, and so on.

In this section, we focus on intra-gender DIFFSVC. We describe following several techniques: 1) derivation of a differential GMM (DIFFGMM), 2) parameter generation algorithm considering global variance (GV), 3) parameter generation algorithm considering modulation spectral (MS), and 4) parameter generation algorithm based on trajectory differential feature.

3.2.1. Training process of the DIFFGMM

For the training process of the DIFFGMM, a joint probability density function of spectral features of the source singer and the differential between the source and target singers is modeled with DIFFGMM, which is directly derived from a traditional GMM[*]. Let $\boldsymbol{D}_t = \left[\boldsymbol{d}_t^\top, \Delta\boldsymbol{d}_t^\top\right]^\top$ denote the static and dynamic differential feature vector, where $\boldsymbol{d}_t = \boldsymbol{y}_t - \boldsymbol{x}_t$. The $2D$-dimensional joint static and dynamic feature vector between the source and the differential features is given by

$$\begin{bmatrix} \boldsymbol{X}_t \\ \boldsymbol{D}_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_t \\ \boldsymbol{Y}_t - \boldsymbol{X}_t \end{bmatrix} = \boldsymbol{A} \begin{bmatrix} \boldsymbol{X}_t \\ \boldsymbol{Y}_t \end{bmatrix}, \tag{3.1}$$

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ -\boldsymbol{I} & \boldsymbol{I} \end{bmatrix}, \tag{3.2}$$

where $\boldsymbol{A}$ is a transformation matrix that transforms the joint feature vector between the source and target features into that of the source and difference features. $\boldsymbol{I}$ denotes the identity matrix. Applying the transformation matrix to the traditional GMM in Equation (2.1), the joint probability density function of the DIFFGMM is derived as follows:

$$P\left(\boldsymbol{X}_t, \boldsymbol{D}_t | \boldsymbol{\lambda}^{(XD)}\right) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix} \boldsymbol{X}_t \\ \boldsymbol{D}_t \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(D)} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XD)} \\ \boldsymbol{\Sigma}_m^{(DX)} & \boldsymbol{\Sigma}_m^{(DD)} \end{bmatrix}\right), \tag{3.3}$$

[*] It is also possible to model the joint probability density function of the DIFFGMM based on expectation-maximization (EM) algorithm using joint feature vector of source and differential.

$$\begin{aligned}
\boldsymbol{\mu}_m^{(D)} &= \boldsymbol{\mu}_m^{(Y)} - \boldsymbol{\mu}_m^{(X)}, && (3.4)\\
\boldsymbol{\Sigma}_m^{(XD)} &= \boldsymbol{\Sigma}_m^{(DX)\top} = \boldsymbol{\Sigma}_m^{(XY)} - \boldsymbol{\Sigma}_m^{(XX)}, && (3.5)\\
\boldsymbol{\Sigma}_m^{(DD)} &= \boldsymbol{\Sigma}_m^{(XX)} + \boldsymbol{\Sigma}_m^{(YY)} - \boldsymbol{\Sigma}_m^{(XY)} - \boldsymbol{\Sigma}_m^{(YX)}. && (3.6)
\end{aligned}$$

3.2.2. Conversion process

For the conversion process, the converted differential feature vector is determined based on the DIFFGMM in the same manner as maximum likelihood parameter generation described in Section 2.3.2. Figure 3.2 (a) indicates the conversion flow of the DIFFSVC based on the DIFFGMM. Time sequence vectors of the source features and the differential features are denoted as $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top]^\top$ and $\boldsymbol{D} = [\boldsymbol{D}_1^\top, \cdots, \boldsymbol{D}_T^\top]^\top$, where T is the number of frames included in the time sequence of the given source feature vectors. A time sequence vector of the converted static features $\hat{\boldsymbol{d}} = [\hat{\boldsymbol{d}}_1^\top, \cdots, \hat{\boldsymbol{d}}_T^\top]^\top$ is determined as follows:

$$\hat{\boldsymbol{d}} = \underset{\boldsymbol{d}}{\mathrm{a}rgmax}\, P\left(\boldsymbol{D}|\boldsymbol{X}, \boldsymbol{\lambda}^{(XD)}\right) \text{ s.t. } \boldsymbol{D} = \boldsymbol{W}\boldsymbol{d}, \qquad (3.7)$$

$$P\left(\boldsymbol{D}|\boldsymbol{X}, \boldsymbol{\lambda}^{(XD)}\right) = \prod_{t=1}^{T} \sum_{m=1}^{M} P\left(m|\boldsymbol{X}_t, \boldsymbol{\lambda}^{(XD)}\right) P\left(\boldsymbol{D}_t|m, \boldsymbol{X}_t, \boldsymbol{\lambda}^{(XD)}\right), \qquad (3.8)$$

$$P\left(\boldsymbol{D}_t|m, \boldsymbol{X}_t, \boldsymbol{\lambda}^{(XD)}\right) = \mathcal{N}\left(\boldsymbol{D}_t; \boldsymbol{E}_{m,t}^{(D)}, \boldsymbol{V}_m^{(D)}\right), \qquad (3.9)$$

$$\begin{aligned}
\boldsymbol{E}_{m,t}^{(D)} &= \boldsymbol{\mu}_m^{(D)} + \boldsymbol{\Sigma}_m^{(DX)} \boldsymbol{\Sigma}_m^{(XX)-1} \left(\boldsymbol{X}_t - \boldsymbol{\mu}_m^{(X)}\right), && (3.10)\\
\boldsymbol{V}_m^{(D)} &= \boldsymbol{\Sigma}_m^{(DD)} - \boldsymbol{\Sigma}_m^{(DX)} \boldsymbol{\Sigma}_m^{(XX)-1} \boldsymbol{\Sigma}_m^{(XD)}. && (3.11)
\end{aligned}$$

Figure 3.3 indicates examples of the spectral envelopes of an original spectral feature, estimated spectral feature differential, and converted spectral feature. In the DIFFSVC method, the original spectral feature is converted into the estimated spectral feature differential feature based on the DIFFGMM. This estimated spectral feature differential varies frame-by-frame. Therefore, time-variant filtering process defined by the estimated spectral feature differential is performed into an original waveform.

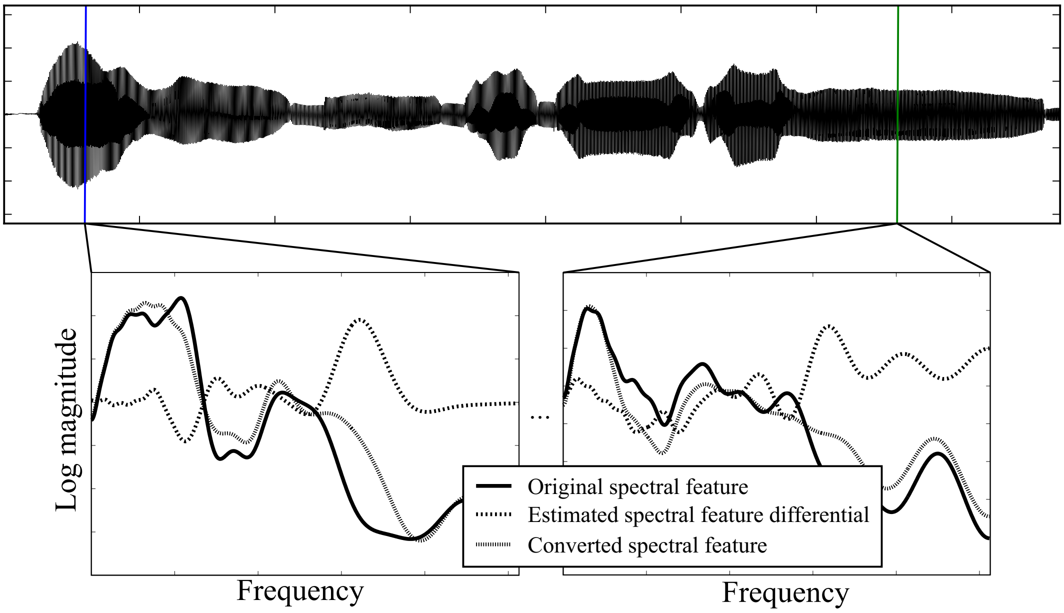


Figure 3.3.: Examples of spectral envelopes of the original spectral feature, estimated spectral feature differential, and converted spectral feature.

Conversion considering GV

In order to alleviate sound quality degradation caused by over-smoothing effect of the converted feature trajectory, we propose parameter generation technique considering GV The converted spectral feature differential trajectory is determined by maximizing a new objective function as follows:

$$\hat{\boldsymbol{d}} = \underset{\boldsymbol{d}}{argmax}\, P\left(\boldsymbol{D}|\boldsymbol{X}, \boldsymbol{\lambda}^{(XD)}\right)^{\omega} P\left(\boldsymbol{v}(\boldsymbol{y}')|\boldsymbol{\lambda}^{(v)}\right) \text{ s.t. } \boldsymbol{D} = \boldsymbol{W}\boldsymbol{d}, \tag{3.12}$$

where $\boldsymbol{y}' = [\boldsymbol{x} + \boldsymbol{d}]$ and the constant ω denotes a parameter for controlling a balance between the two likelihood functions. The converted feature differential trajectory is iteratively updated by using the steepest descent method as bellow:

$$\hat{\boldsymbol{d}}^{(i+1)-th} = \hat{\boldsymbol{d}}^{(i)-th} + \alpha \cdot \Delta\hat{\boldsymbol{d}}^{(i)-th}, \tag{3.13}$$

where α is a step size parameter. The gradient vector $\Delta\hat{\boldsymbol{d}}^{(i)-th}$ is given by

$$\Delta\boldsymbol{d}^{(i)-th} = \left.\frac{\partial\mathcal{L}}{\partial\boldsymbol{d}}\right|_{\boldsymbol{d}=\boldsymbol{d}^{(i)-th}}, \tag{3.14}$$

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\boldsymbol{d}} &= \omega\left(-\boldsymbol{W}^\top\boldsymbol{V}_m^{(D)^{-1}}\boldsymbol{W}\boldsymbol{d} + \boldsymbol{W}^\top\boldsymbol{V}_m^{(D)^{-1}}\boldsymbol{E}_m^{(D)}\right) \\
&\qquad + \left[\boldsymbol{v}_1'^\top, \boldsymbol{v}_2'^\top, \cdots, \boldsymbol{v}_t'^\top, \cdots, \boldsymbol{v}_T'^\top\right]^\top, && (3.15) \\
\boldsymbol{E}_m^{(D)} &= \left[\boldsymbol{E}_{m_1,1}^{(D)}, \cdots, \boldsymbol{E}_{m_t,t}^{(D)}, \cdots, \boldsymbol{E}_{m_T,T}^{(D)}\right]^\top, && (3.16) \\
\boldsymbol{V}_m^{(D)^{-1}} &= \mathrm{diag}\left[\boldsymbol{V}_{m_1}^{(D)^{-1}}, \cdots, \boldsymbol{V}_{m_t}^{(D)^{-1}}, \cdots, \boldsymbol{V}_{m_T}^{(D)^{-1}}\right], && (3.17) \\
\boldsymbol{v}_t' &= \left[v_t'(1), v_t'(2), \cdots, v_t'(d), \cdots, v_t'(D)\right]^\top, && (3.18) \\
v_t'(d) &= -\frac{2}{T}\boldsymbol{p}^{(v)}(d)^\top\left(\boldsymbol{v}(\boldsymbol{y}') - \boldsymbol{\mu}^{(v)}\right)\left(y_t'(d) - \bar{y}'(d)\right), && (3.19)
\end{aligned}$$

where $\boldsymbol{p}^{(v)}(d)$ indicates a d-th column vector of the inverse matrix of $\boldsymbol{\Sigma}^{(vv)}$. An initial differential feature vector sequence for the iterative update is determined as follows:

$$\hat{d}_t'(d) = \sqrt{\frac{\mu_v(d)}{v(d)}}\left(\hat{y}_t(d) - \bar{\hat{y}}(d)\right) + \bar{\hat{y}}(d) - x_t(d), \tag{3.20}$$

where $\hat{y}_t(d)$ indicates the converted differential feature vector at frame t determined by the DIFFSVC and $\bar{\hat{y}}(d)$ indicates its average over a time sequence.

It has been reported that unvoiced consonants (e.g. /s/, /sh/) are less affected by speaker individuality compared with voiced sounds (e.g. /ae/, /n/) in normal speech [47]. Based on this finding, in order to alleviate the over-smoothing effect as much as possible, we minimize the amount of conversion at unvoiced frames by smoothing the converted feature differential at those frames. We implement this process on top of the previously described DIFFSVC with GV by modifying $\boldsymbol{E}_{m,t}^{(D)}$ and $\boldsymbol{V}_m^{(D)^{-1}}$ at unvoiced frames as follows:

$$\boldsymbol{E}_{m,t}^{(D)} = \left\{ \boldsymbol{0} \quad \text{(for static \& delta)}, \right. \tag{3.21}$$

$$\boldsymbol{V}_m^{(D)^{-1}} = \begin{cases} \boldsymbol{0} & \text{(for static)} \\ \boldsymbol{V}_m^{(\Delta D)^{-1}} & \text{(for delta)}, \end{cases} \tag{3.22}$$

where $\boldsymbol{V}_m^{(\Delta D)^{-1}}$ shows delta components of the inverse matrix of the covariance matrix in Equation (3.9). These parameter modifications make the converted spectral feature differential smoothly varies at unvoiced frames. Note that we avoid updating the converted spectral feature differential at the unvoiced frames in Equation (3.13).

Post-filtering considering MS

Statistical modeling tends to deteriorate modulation components of the converted parameters even when considering GV, and keeping natural MSs is strongly effective for improving the quality of the converted voice. An MS-based post-filter [11], which is applied after parameter generation in traditional VC based on the GMM, modifies a converted parameter sequence so that the sequence has the target singer's natural MS. Here, we propose an MS-based post-filtering process that modifies spectral differentials, $\hat{\boldsymbol{d}}$, such that the finally synthesized has the target singer's natural MS.

In training, we calculate MS statistics for target singer's natural and converted parameters from the training data, $\boldsymbol{y}$ and $\tilde{\boldsymbol{y}} = [\hat{\boldsymbol{d}} + \boldsymbol{x}^{(\mathrm{L}PF)}]$ where $\boldsymbol{x}^{(\mathrm{L}PF)}$ denotes smoothed spectral feature sequence of source singing voice. Here, let $\mu_{d,f}^{(y)}$ and $\mu_{d,f}^{(\tilde{y})}$ be the mean of the f-th MS of d-th dimension target parameter $s_{d,f}(\boldsymbol{y})$ and MS of converted parameter $s_{d,f}(\tilde{\boldsymbol{y}})$, and let $\sigma_{d,f}^{(y)}$ and $\sigma_{d,f}^{(\tilde{y})}$ be their variance. The $\hat{\boldsymbol{d}}$ is generated by converting $\boldsymbol{x}^{(\mathrm{L}PF)}$.

For conversion process, $\boldsymbol{x}^{(\mathrm{L}PF)}$ is first added to the generated $\hat{\boldsymbol{d}}$. Then, the MS, $s_{d,f}(\tilde{y})$ is converted as follows:

$$s'_{d,f}(\tilde{y}) = \frac{\sigma_{d,f}^{(y)}}{\sigma_{d,f}^{(\tilde{y})}} \left(s_{d,f}(\tilde{y}) - \mu_{d,f}^{(\tilde{y})} \right) + \mu_{d,f}^{(y)}. \tag{3.23}$$

The converted $\tilde{\boldsymbol{y}}$ is determined using the converted MS and the original phase components. The MSPFed spectral differentials, $\hat{\boldsymbol{d}}^{(\mathrm{MS}PF)}$ can be determined by subtracting $\boldsymbol{x}^{(\mathrm{L}PF)}$ from the converted $\tilde{\boldsymbol{y}}$ [†]. Note that, in this thesis, we use mean-normalized MSs and adopt a segment-level post-filtering process [11].

[†] Note that, because the MSPF process is non-linear to the parameter sequence, the sequence that $\boldsymbol{x}^{(\mathrm{L}PF)}$ is subtracted from the converted $\tilde{\boldsymbol{y}}$ is not equal to $\hat{\boldsymbol{d}}$.

Conversion based on trajectory differential feature

In the DIFFSVC based on the DIFFGMM, the spectral feature differential is estimated based on the joint probability density function of the DIFFGMM in joint static and dynamic feature space described in Equation (3.3). Against this differential parameter generation technique, we propose another parameter generation technique of the spectral differential based on a probability density function in static feature trajectory space obtained the knowledge of trajectory model [48] in statistical parametric speech synthesis. Figure 3.2 (b) indicates the conversion process of the DIFFSVC based on trajectory differential feature. In order to model the probability density function of the spectral differential in static feature space, in this technique, two probability density functions of the traditional GMM in Equation (2.1) and the intra-singer GMM in Equation (4.1) are modeled in advance. The traditional GMM converts the acoustic feature of the source singer into that of the target singer, whereas the intra-singer GMM makes it possible to convert the acoustic feature of the source singer into averaged acoustic feature of the source singer.

For the traditional GMM, the probability density function of the static feature vector is derived by approximating conditional probability density function $P\left(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}^{(XY)}\right)$ based on sub-optimum mixture component sequences $\hat{\boldsymbol{m}} = [\hat{m}_1, \cdots, \hat{m}_T]$. The probability density function is given by

$$P\left(\boldsymbol{y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}^{(XY)}\right) = \mathcal{N}\left(\boldsymbol{y}; \hat{\boldsymbol{y}}_{\hat{\boldsymbol{m}}}, \boldsymbol{P}_m\right) \tag{3.24}$$

$$\hat{\boldsymbol{y}}_{\hat{\boldsymbol{m}}} = \underset{\boldsymbol{y}}{a rgmax}\, P\left(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}^{(XY)}\right) \text{ s.t. } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \tag{3.25}$$

For the intra-singer GMM, the probability density function of the static feature vector is derived using intra-singer GMM in the same manner as the traditional GMM. The probability density function is given by

$$P\left(\boldsymbol{x}'|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}^{(XYX)}\right) = \mathcal{N}\left(\boldsymbol{x}'; \hat{\boldsymbol{x}}'_{\hat{\boldsymbol{m}}}, \boldsymbol{Q}_m\right) \tag{3.26}$$

$$\hat{\boldsymbol{x}}'_{\hat{\boldsymbol{m}}} = \underset{\boldsymbol{x}'}{a rgmax}\, P\left(\boldsymbol{X}'|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}^{(XYX)}\right) \text{ s.t. } \boldsymbol{X}' = \boldsymbol{W}\boldsymbol{x}' \tag{3.27}$$

where $\boldsymbol{\lambda}^{(XYX)}$ denotes a parameter set of the intra-singer GMM.

The probability density function of the feature vector differential in static feature space $\boldsymbol{d} = [\boldsymbol{y} - \boldsymbol{x}']$ is given by

$$P\left(\boldsymbol{d}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}^{(xd)}\right) = \mathcal{N}\left(\boldsymbol{d}; \hat{\boldsymbol{y}}_{\hat{\boldsymbol{m}}} - \hat{\boldsymbol{x}}_{\hat{\boldsymbol{m}}}, \boldsymbol{P}_m + \boldsymbol{Q}_m\right), \tag{3.28}$$

where $\boldsymbol{\lambda}^{(xd)}$ denotes a parameter set of Gaussian distribution of the spectral feature differential in static feature space. Consequently, the static spectral feature trajectory is estimated as maximum likelihood estimate of the probability density function.

$$\begin{aligned}\hat{\boldsymbol{d}} &= \underset{\boldsymbol{d}}{\mathrm{a}rgmax}\, P\left(\boldsymbol{d}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}^{(xd)}\right) \\ &= [\hat{\boldsymbol{y}}_{\hat{\boldsymbol{m}}} - \hat{\boldsymbol{x}}'_{\hat{\boldsymbol{m}}}].\end{aligned} \tag{3.29}$$

3.3. Intra/inter-gender DIFFVC using F_0 transformation

We propose several F_0 transformation techniques for intra/inter-gender DIFFVC to make it possible to widely accept various F_0 transformation ratios. In the VC for normal speech, the F_0 transformation ratios vary depending on a combination of source and target speakers although the F_0 transformation ratios of inter-gender singing VC are usually twice or half. In order to apply the statistical waveform modification technique into not only singing voice but also normal speech, it is necessary to implement VC with continuous F_0 transformation ratios. In this section, in order to achieve VC with continuous F_0 transformation ratio, following F_0 transformation techniques using with or without the vocoding process are proposed: 1) DIFFVC with F_0 transformation using STRAIGHT vocoder, 2) DIFFVC with F_0 transformation based on the residual signal modification using time-scaling and resampling, and 3) DIFFVC with F_0 transformation based on waveform modification using time-scaling and resampling.

3.3.1. F_0 transformation using STRAIGHT

Figure 3.4 describes the conversion process of the DIFFVC method with the F_0 transformation based on STRAIGHT vocoder. In this method, several acoustic features

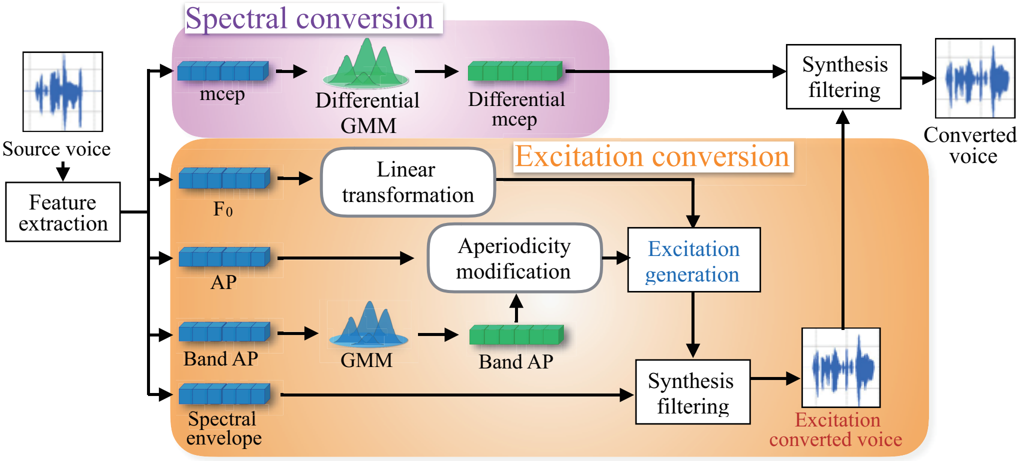


Figure 3.4.: Conversion process of DIFFVC w/ F_0 transformation using STRAIGHT vocoder.

such as F_0, aperiodicity, and spectral envelope are extracted from the source voice using STRAIGHT analysis framework [49]. For the excitation conversion, F_0 is transformed based on global linear transformation in the same manner as the traditional VC method. The aperiodic components at all frequency bins are shifted using band-averaged aperiodic differentials between the extracted and converted ones as a global bias term. Then, an F_0 transformed source voice is synthesized using a full representation of STRAIGHT spectral envelope, the transformed F_0, and the transformed aperiodic components. Finally, spectral envelope of the F_0 transformed source voice is converted using the converted mel-cepstrum differentials with DIFFGMM in the same manner as the DIFFSVC.

This method is capable of converting the excitation parameters including not only F_0 but also aperiodic components as accurately as in the conventional VC. Therefore, it is expected that the conversion accuracy of speaker identity is almost equivalent to that of the conventional VC. On the other hand, this method ruins the advantage of the DIFFVC method, i.e., achievement of a high-quality converted voice by avoiding the vocoding process. Consequently, this method significantly suffers from quality degradation of the converted voice caused by F_0 extraction errors, unvoiced/voiced decision errors, lack of natural phase components, and so on.

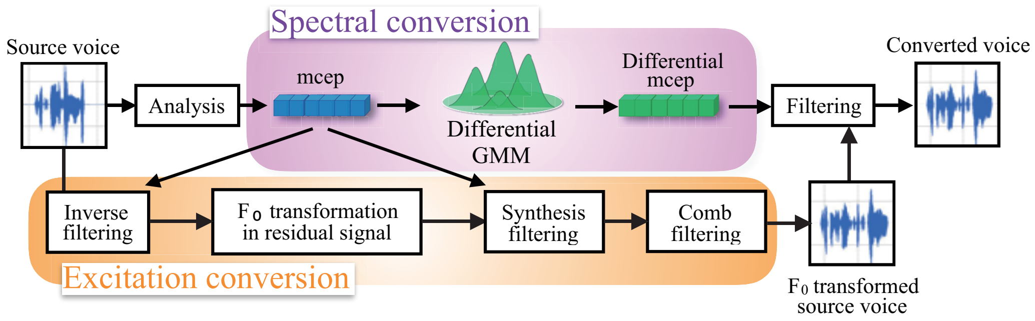


Figure 3.5.: Conversion process of DIFFVC w/ F_0 transformation by residual signal modification.

3.3.2. F_0 transformation by residual signal modification

Figure 3.5 describes the conversion process of the DIFFVC method with F_0 transformation based on residual signal modification. In this method, the F_0 transformation is carried out by directly modifying the residual signal. For the excitation conversion, the residual signal composed of harmonic and aperiodic components is extracted from the source voice with inverse filtering based on the extracted mel-cepstrum. Then, the time-scaling with waveform similarity based overlap-add (WSOLA) [50] and resampling is performed on the residual signal in order to transform F_0. For instance, if F_0 is transformed to higher, the residual signal is expanded to make its duration longer, followed by using down-sampling to restore the length of the residual signal. If F_0 is transformed to lower, the residual signal is shrunk to make its duration shorter, followed by using up-sampling to restore its length. We further need to perform an additional process when decreasing F_0, making high-frequency components of the transformed residual signal vanish. To reconstruct these vanished frequency components, they are generated using a noise excitation signal because the high-frequency components of a speech signal tend to be less periodic and be well modeled with noise components. The F_0 transformed source voice is generated by filtering the resulting residual signal again using the extracted mel-cepstrum. Finally, spectral envelope of the F_0 transformed source voice is converted using the converted mel-cepstrum differentials with DIFFGMM in the same manner as the DIFFSVC. Note that we set the F_0 transformation ratio to a constant value for each speaker pair.

In this technique, a part of natural phase components of the source voice is well

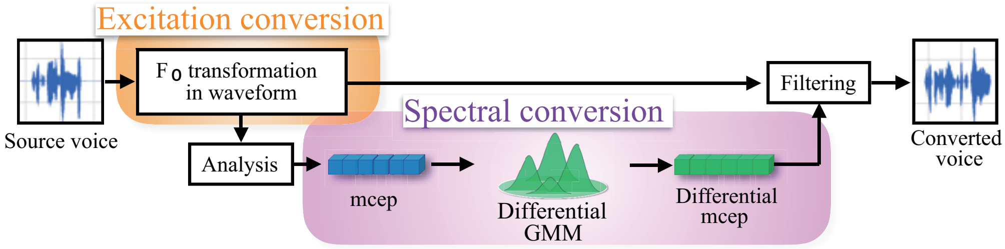


Figure 3.6.: Conversion process of DIFFVC w/ F_0 transformation by waveform modification.

preserved because the F_0 transformation is performed by directly modifying the residual signal without the vocoding process. Moreover, this technique makes it possible to freely control the F_0 transformation ratio without changing DIFFGMM for the spectral differential conversion because the original spectral envelope is also preserved through the F_0 transformation. On the other hand, it is possible to cause speech quality degradation of the converted voice due to some essentially difficult processes, e.g., the difficulty of extracting the residual signal by perfectly removing the effect of the spectral envelope.

3.3.3. F_0 transformation by waveform modification

Figure 3.6 illustrates the conversion process of the DIFFVC method with the F_0 transformation using waveform modification. In this method, the F_0 transformation using WSOLA and resampling based on linear interpolation is directly applied to an original waveform of the source voice. Because this direct waveform modification causes frequency warping, the spectral envelope also changes according to the F_0 transformation ratio. Therefore, we need to use DIFFGMM capable of converting such a frequency warped source voice. We train the joint GMM using the F_0 transformed source voices and the natural target voices. For spectral conversion, the converted voice is generated by filtering the F_0 transformed source voice with converted mel-cepstrum differential determined with DIFFGMM derived from the corresponding joint GMM. The F_0 transformation ratio is set to a constant value for each speaker pair. Note that this F_0 transformation doesn't cause any problems even when decreasing F_0 because the high-frequency components are generated with aliasing caused by the linear interpolation and the resulting spectral envelope is modeled with the joint GMM and also DIFFGMM.

In this technique, there is no approximation error caused by the vocoding process and the other processes, such as inverse filtering. Therefore, it is expected that this method achieves high-quality of the converted voice. Moreover, this method is based on quite simple processes, and therefore, it is easy to implement it to the real-time VC system [25]. On the other hand, we need to separately train the joint GMM for each different setting of the F_0 transformation ratio because spectral envelope of the F_0 transformed source voice depends on the F_0 transformation ratio.

3.4. Experimental evaluation

3.4.1. Evaluation of intra-gender DIFFSVC

Experimental condition

In this evaluation, we denoted several conventional SVC and proposed DIFFSVC techniques as follow:

SVC (w/o GV)
The conventional SVC based on the GMM method w/o considering GV

SVC (w/ GV)
The conventional SVC based on the GMM method w/ considering GV

DIFFSVC (w/o GV)
The proposed DIFFSVC method based on the DIFFGMM w/o considering GV

DIFFSVC (w/ GV)
The proposed DIFFSVC method based on the DIFFGMM w/ considering GV

TrjDiff
The proposed DIFFSVC method based on trajectory differential spectral feature

We used singing voices of 21 Japanese traditional songs, which were divided into 152 phrases, where the duration of each phrase was approximately 8 seconds. Amateur singers including 3 males and 3 females sang these phrases. The sampling frequency was set to 16 kHz. STRAIGHT [12] was used to extract spectral envelopes, which were parameterized to the 1-24th, 1-32th, and 1-40th mel-cepstral coefficients as spectral features. As the source excitation features for the conventional SVC method, we used F_0 and aperiodic components in five frequency bands, i.e., 0-1, 1-2, 2-4, 4-6, and 6-8 kHz, which were also extracted by STRAIGHT [49]. The frame shift was 5 ms. The mel log spectrum approximation (MLSA) filter [40] was used as the synthesis filter in both the conventional SVC and the proposed DIFFSVC methods. We used 80 phrases for the GMM training and the remaining 72 phrases were used for evaluation. The speaker-dependent GMMs were separately trained for individual singer pairs determined in a round-robin fashion within intra-gender singers. The number of mixture components for the mel-cepstral coefficients was 128 and for the aperiodic components was 64.

Objective evaluation

As an objective evaluation, we compared the mel-cepstral distortion of the converted feature trajectories. The mel-cepstral distortion (Mel-CD) is calculated as

$$\text{Mel-CD } [dB] = \frac{10}{\ln 10} \sqrt{2 \sum_{d=1}^{24} \left(mc_d^{(X)} - mc_d^{(Y)} \right)^2}, \tag{3.30}$$

where $mc_d^{(X)}$ and $mc_d^{(Y)}$ represent the d^{th} dimensional component of the converted mel-cepstrum and that of the target mel-cepstrum, respectively. The number of the order of mel-cepstrum was set to 24. For the proposed DIFFSVC methods, converted mel-cepstrum was extracted from the converted singing voice using STRAIGHT. In order to evaluate the effectiveness of the DIFFGMM modeling technique based on directly parameter transformation from the traditional GMM described in Section 3.2.1, we also evaluated the Mel-CD of the DIFFGMM modeling technique using joint source and differential feature vector as a reference. We denote DIFFSVC xdj (w/o GV) and DIFFSVC xdj (w/ GV) to the DIFFSVC based on the DIFFGMM directly modeled based on EM algorithm using joint feature vector of source and differential.

Table 3.1 indicates the experimental results of the Mel-CD between the mel-cepstrum extracted from source singer's natural singing voice and converted mel-cepstrum, and between the mel-cepstrum extracted from target singer's natural singing voice and con-

Table 3.1.: Mel-cepstral distortions of several conversion methods.

	Mel-CD [dB]	
Method	Source singer	Target singer
SVC (w/o GV)	6.09	5.40
SVC (w/ GV)	6.70	6.01
DIFFSVC (w/o GV)	5.80	5.21
DIFFSVC xdj (w/o GV)	5.79	5.21
DIFFSVC (w/ GV)	6.20	5.73
DIFFSVC xdj (w/ GV)	6.23	5.77
TrjDiff	4.73	5.24

verted mel-cepstrum. We can see that there is a tendency to increase the Mel-CD when considering GV in SVC, DIFFSVC, and DIFFSVC xdj methods. It is known that gain of Mel-CD due to considering GV does not affect any sound quality degradation of the converted voice in VC for normal speech [10]. The proposed DIFFSVC methods tend to be achieved smaller Mel-CDs regard to the source singer compared with those of the conventional SVC methods. Therefore, it is expected that the sound quality of the proposed DIFFSVC methods is close to the natural singing voice of the source singer compared with those of the conventional SVC methods. As for the Mel-CDs regard to the target singer, there is a small difference between the proposed DIFFSVC methods and the conventional SVC methods. This implies that the remaining components of the source singer after conversion may increase the Mel-CD in the proposed DIFFSVC methods. In terms of the TrjDiff method, it is considered that the converted voice closes to not the target singer but the source singer because the Mel-CD for the target singer is larger than the Mel-CD for the source singer. For the DIFFGMM modeling technique, it is considered that the parameter transformation technique does not have any bad effect on the sound quality because there are no significant differences between the Mel-CDs of the DIFFSVC and DIFFSVC xdj methods.

From these results, we can say following things, 1) the use of the GV makes higher Mel-CD for not only the conventional SVC methods but also the proposed DIFFSVC methods, 2) there are small differences of the conversion accuracy between the SVC and DIFFSVC methods, on the other hand, the DIFFSVC methods slightly remain the components of the source mel-cepstrum compared with those of SVC methods,

3) The TrjDiff method makes it possible to bet smaller Mel-CD regard to the target singer compared with the other methods, but the Mel-CD regard to the source singer is smaller, and 4) there is no significant difference between DIFFGMM modeling techniques.

Subjective evaluation

We subjectively evaluated sound quality and singer identity of the converted singing voices to compare the conventional SVC and the proposed DIFFSVC methods. In the subjective evaluation, four preference tests were performed.

The first and second preference tests evaluated to compare the conventional SVC (w/ GV) and proposed DIFFSVC (w/o GV) methods. The first preference test evaluated sound quality of the converted singing voices of the SVC (w/ GV) and DIFFSVC (w/o GV) methods. The converted singing voice samples of the SVC (w/ GV) and the DIFFSVC (w/o GV) methods for the same phrase were presented to subjects in random order. The subjects selected which sample had better sound quality. The second preference test evaluated the conversion accuracy on singer identity of the converted singing voices for SVC (w/ GV) and DIFFSVC (w/o GV) methods. A natural singing voice sample of the target singer was presented to the subjects first as a reference. Then, the converted singing voice samples of the SVC (w/ GV) and the DIFFSVC (w/o GV) methods for the same phrase were presented in random order. The subjects selected which sample was more similar to the reference natural singing voice in terms of singer identity. We varied the order settings of the mel-cepstral coefficients to confirm the effects of higher order of mel-cepstral coefficients. The number of subjects in the first and second evaluation was 8 and each listener evaluated 24 sample pairs in each order setting of the mel-cepstral coefficients. All subjects don't specialize in audio. Subjects were allowed to replay each sample pair as many times as necessary.

Figure 3.7 indicates the results of the preference test between SVC (w/ GV) and DIFFSVC (w/o GV) methods for the sound quality. The DIFFSVC (w/o GV) method makes it possible to generate the converted speech with better sound quality than the SVC (w/ GV) in any order setting of the mel-cepstral coefficients. This is assumed that the DIFFSVC (w/o GV) is free from various errors caused by the waveform generation based on the vocoding, such as F_0 extraction errors or spectral modeling errors caused by liftering. And, we can see that the differential of the order setting of mel-cepstral

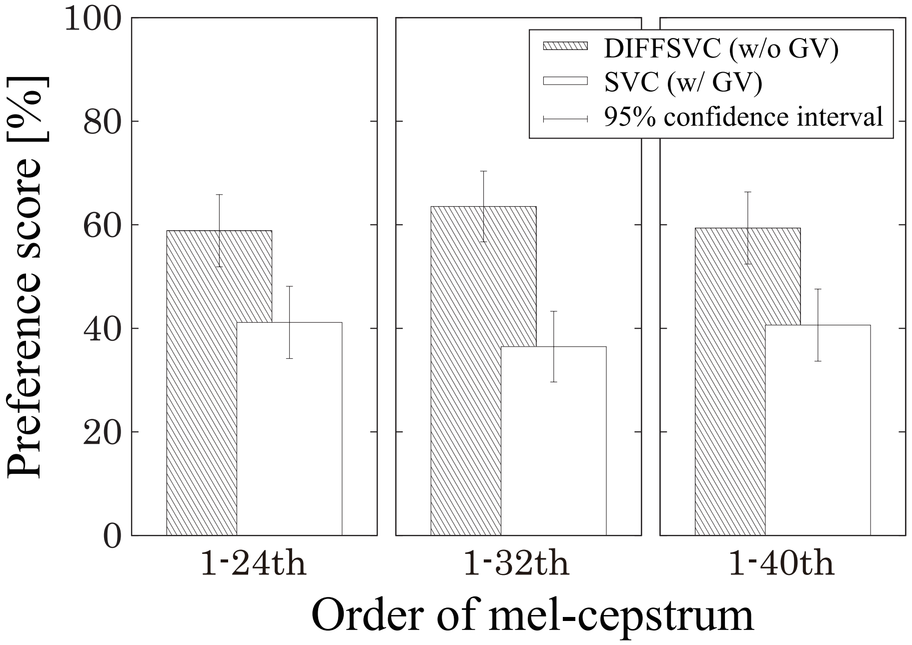


Figure 3.7.: Evaluation of sound quality for SVC (w/ GV) and DIFFSVC (w/o GV) methods.

coefficients has little effect.

Figure 3.8 indicates the results of the preference test between SVC (w/ GV) and DIFFSVC (w/o GV) for the singer identity. The conversion accuracy of the singer identity of the DIFFSVC (w/o GV) is not statistically significantly different from that of the SVC (w/ GV) in any order setting of the mel-cepstral coefficients. This result suggests that the aperiodic components have little effect on the singer identity in singing voices, and even if the DIFFSVC (w/o GV) cannot convert the excitation features, the conversion accuracy of the singer identity still remains equivalent to that of the SVC (w/ GV).

These two results demonstrate that the DIFFSVC (w/o GV) method is capable of converting the voice timbre with higher sound quality while causing no degradation in the conversion accuracy of singer identity compared with the conventional SVC. From this result, we only set the order settings of the mel-cepstrum to 24 in following evaluations.

The third and fourth preference tests evaluated the effectiveness several parameter

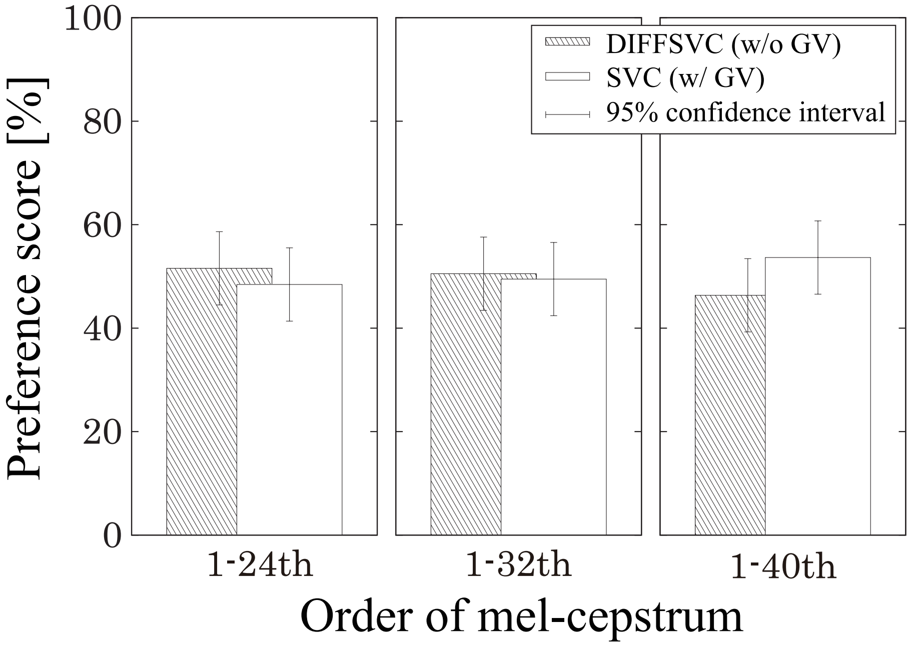


Figure 3.8.: Evaluation of conversion accuracy on singer individuality for SVC (w/ GV) and DIFFSVC (w/o GV) methods.

generation techniques by comparing with DIFFSVC (w/o GV), DIFFSVC (w/ GV), and TrjDiff methods. The third preference test evaluated sound quality of the converted singing voices of DIFFSVC (w/o GV), DIFFSVC (w/ GV), and TrjDiff methods. The two converted singing voice samples of the same phrase for comparisons were presented to subjects in random order. The subjects selected which sample had better in terms of sound quality. The fourth preference test evaluated the singer identity conversion accuracy for comparisons. A natural singing voice sample of the target singer was presented to the subjects first as a reference. Then, the two converted singing voice samples of the DIFFSVC (w/o GV), DIFFSVC (w/ GV), and TrjDiff methods for the same phrase were presented to subjects in random order. The subjects selected which sample was more similar in the same manner as the second preference test. The number of subjects was 6 and each listener evaluated 54 sample pairs in the third and forth preference tests. All subjects don't specialize in audio. Subjects were allowed to replay each sample pair as many times as necessary.

Figure 3.9 indicates the result of the preference test for the sound quality between

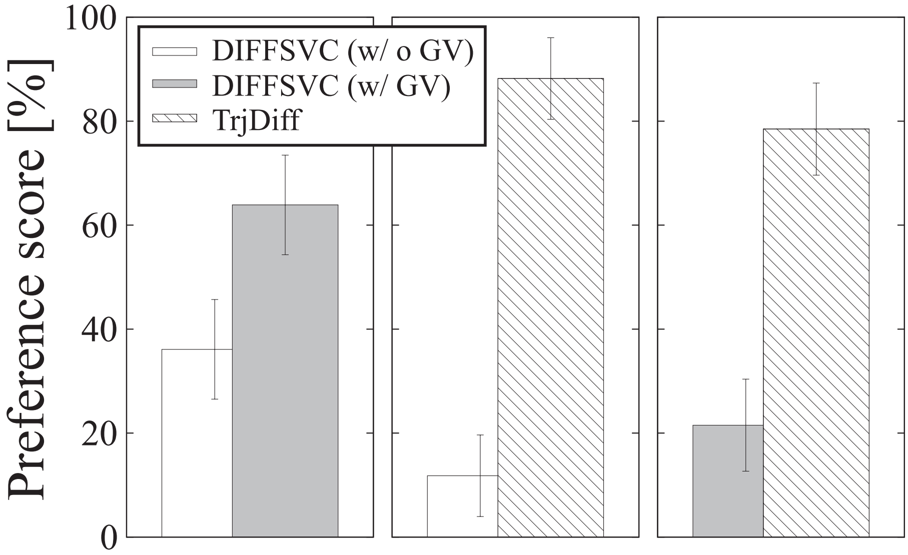


Figure 3.9.: Evaluation of sound quality for DIFFSVC (w/o GV), DIFFSVC (w/ GV), and TrjDiff methods.

DIFFSVC (w/o GV), DIFFSVC (w/ GV), and TrjDiff methods. The DIFFSVC (w/ GV) and TrjDiff methods generate the converted speech with better sound quality than the DIFFSVC (w/o GV) method. As for the experimental result between DIFFSVC (w/o GV) and DIFFSVC (w/ GV) methods, we can see that the parameter generation considering GV is effective on the sound quality not only in the conventional SVC method [10] but also in the proposed DIFFSVC method. As for the experimental results of the TrjDiff method, the TrjDiff method makes it possible to significantly improve the sound quality compared with the other parameter generation methods.

Figure 3.10 indicates the result of the preference test for the singer identity between DIFFSVC (w/o GV), DIFFSVC (w/ GV), and TrjDiff methods. The conversion accuracy on the singer identity of the DIFFSVC (w/ GV) method is not significantly different from that of the DIFFSVC (w/o GV) method. Although the proposed DIFFSVC (w/ GV) method avoids accurately converting spectral features at unvoiced frames, it still yields conversion accuracy of singer individuality almost equal to that of the DIFFSVC (w/o GV) method. On the other hand, the TrjDiff method has a degradation of conversion accuracy compared with that of DIFFSVC (w/o GV).

These two results demonstrate that the DIFFSVC (w/ GV) method is capable of

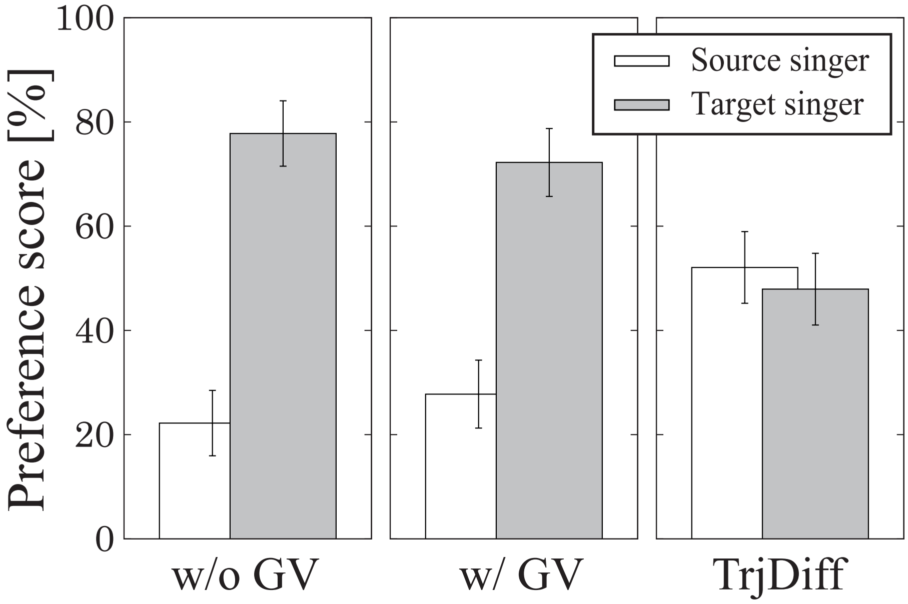


Figure 3.10.: Evaluation of conversion accuracy on singer individuality for DIFFSVC (w/o GV), DIFFSVC (w/ GV), and TrjDiff methods.

converting voice timbre with higher sound quality while causing no degradation in the conversion accuracy of singer identity compared with the DIFFSVC (w/o GV) method. And, although the conversion accuracy of singer individuality is slightly decreasing, the TrjDiff method makes it possible to convert with the significantly higher sound quality compared with the other parameter generation methods.

Analysis of converted feature trajectories

To more deeply analyze what yields naturalness improvements in the proposed DIFFSVC methods, we investigated the difference of the conventional SVC methods and the proposed DIFFSVC methods. We denote several estimated and converted spectral feature follows:

Source

mel-cepstral coefficients extracted from the source singer's natural singing voice

Target

mel-cepstral coefficients extracted from the target singer's natural singing voice

DIFFSVC w/o GV (estimated)

differences of mel-cepstral coefficients estimated with the differential GMM in the DIFFSVC method without considering GV

DIFFSVC w/o GV (converted)

mel-cepstral coefficients extracted from the singing voice converted in the DIFFSVC (w/o GV) method without considering GV

DIFFSVC w/ GV (estimated)

differences of mel-cepstral coefficients estimated with the differential GMM in the DIFFSVC method w/ considering GV

DIFFSVC w/ GV (converted)

mel-cepstral coefficients extracted from the singing voice converted in the DIFFSVC method w/ considering GV

TrjDiff (estimated)

differences of mel-cepstral coefficients estimated with the intra- and inter-GMMs in the DIFFSVC method based on the trajectory differential feature

TrjDiff (converted)

mel-cepstral coefficients extracted from the singing voice converted in the DIFFSVC method based on the trajectory differential feature

Figure 5.5 shows the GVs calculated from several trajectories of mel-cepstral coefficients. The GV in the DIFFSVC w/o GV (converted) significantly decreases compared with that of Target singer. On the other hand, the GV in DIFFSVC w/ GV (converted) and TrjDiff (converted) is very close to that of Target singer. As reported in the previous work [10], the GVs of the converted mel-cepstral coefficients tend to be smaller in SVC w/o GV and this tendency is clearly observed especially in higher-order mel-cepstral coefficients. And, by considering GV in parameter generation of SVC, the GV of the converted trajectories of SVC is almost equivalent to those of the target Target singer. This GV restoration yields significant improvements in sound quality of the converted singing voice. These tendencies are clearly observed in the converted trajectories of the DIFFSVC (w/o GV) and DIFFSVC (w/ GV). Moreover, the GV of the feature differential trajectories in DIFFSVC w/ GV (estimated) is still similar to

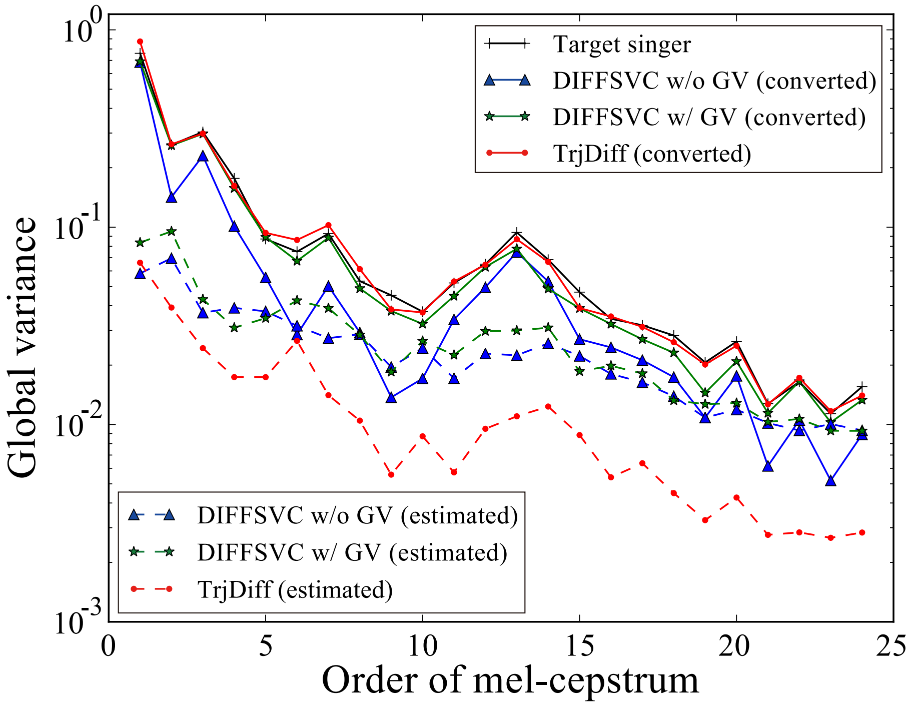


Figure 3.11.: GVs of several mel-cepstral sequences.

those of DIFFSVC w/o GV (estimated). On the other hand, the GV of the feature differential trajectories in TrjDiff (estimated) is quite smaller than the others because the estimated trajectory of TrjDiff (estimated) is the difference between over-smoothed (converted) target spectral by SVC and source spectral by intra-singer SVC. This result implies that the DIFFSVC based on the trajectory differential feature does not restore the GV of converted feature trajectory but cleverly utilize the GV of the natural spectral feature trajectory. These results show the effectiveness of the proposed DIFFSVC methods does not model the GV of the differential trajectory but the GV of the converted trajectory.

Figure 3.12 shows trajectories of the mel-cepstral coefficients and logarithmic F_0 trajectories in each sample. It can be observed from Source and Target that higher-order mel-cepstral coefficients tend to have rapidly varying fluctuations. It has been reported in [42] that these fluctuations are well modeled by the modulation spectrum and strongly affect the sound quality of the converted speech. In the DIFFSVC w/ GV (estimated), the converted feature differential trajectory is smoothly connected from the

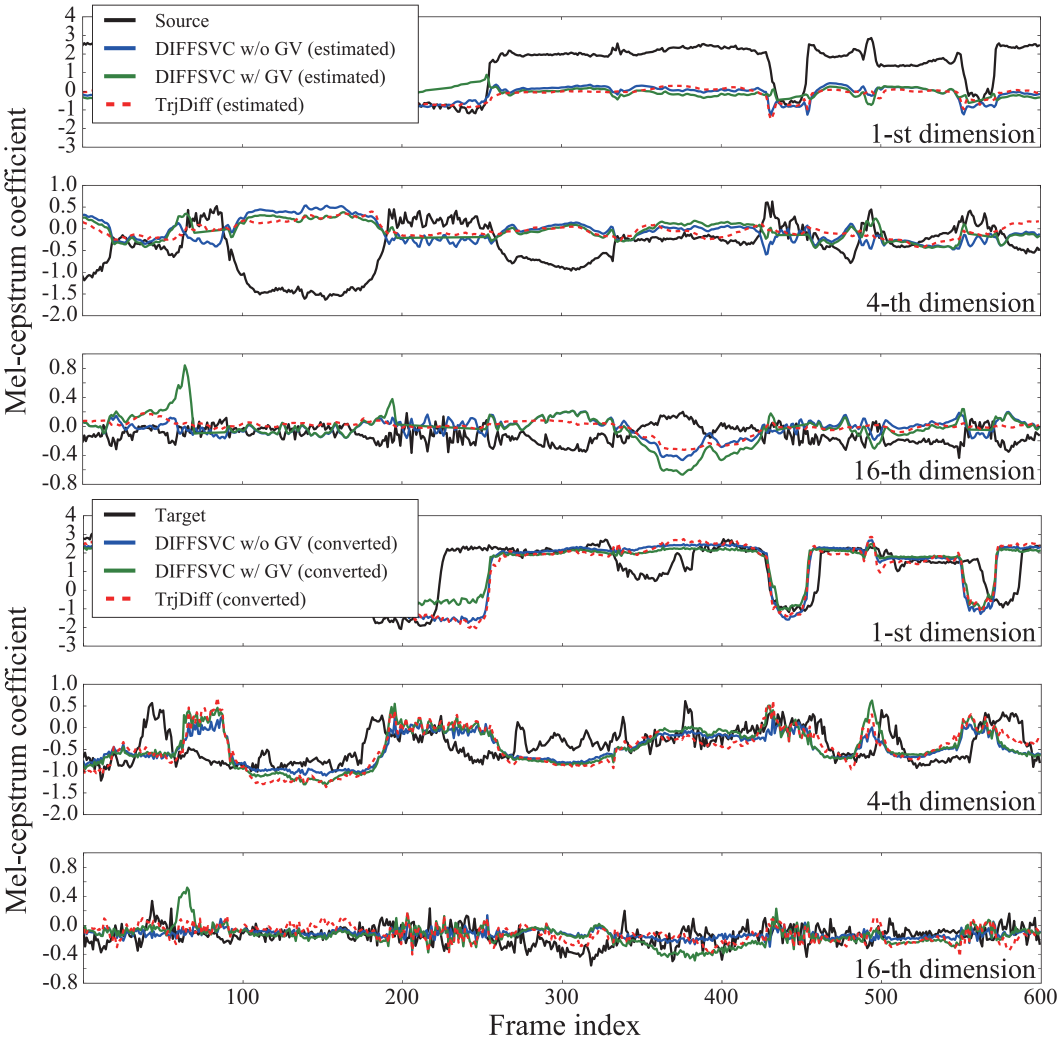


Figure 3.12.: Example of trajectories of spectral feature sequences. Note that the duration of Target trajectories is different from the other trajectories.

end of voiced segments to the start of voiced frames thanks to the smoothing process at unvoiced frames in Sect. 3.2.2. This yields a converted feature trajectory DIFFSVC w/ GV (converted) maintaining natural fluctuations at unvoiced frames. On the other hand, these fluctuations are obviously reduced in the DIFFSVC w/o GV (converted). We can also see that the GV of the converted feature trajectory at higher-order mel-cepstral coefficients is restored more effectively by the DIFFSVC w/ GV (converted) compared with the DIFFSVC w/o GV (converted). Additionally, it can be seen that the trajectory of the TrjDiff (estimated) is much smoother than the others and it affects the converted trajectory in TrjDiff (converted) to maintain the modulation component compared with the other converted feature trajectory. These results imply that the proposed DIFFSVC methods effectively approximate the target spectral fluctuations by using those of the source spectral trajectory and the GV of the target spectral trajectory.

3.4.2. Evaluation of intra/inter-gender DIFFVC

In this section, we evaluate performance of the following DIFFVC methods using F_0 transformation techniques as follows:

- DIFFVC w/ STRAIGHT: The DIFFVC method with F_0 transformation using STRAIGHT vocoding described in Section 3.3.1,
- DIFFVC w/ RES: The DIFFVC method with F_0 transformation based on the residual signal modification [51] described in Section 3.3.2,
- DIFFVC w/ WAV: The DIFFVC method with F_0 transformation based on the waveform modification described in Section 3.3.3.

Experimental condition

We evaluated sound quality and speaker identity of the converted voices to compare the performance of the different F_0 transformation techniques in both intra-gender and cross-gender conversions tasks. We used the English speech database used in the Voice Conversion Challenge (VCC) 2016 [52]. The number of evaluation speakers was 10 consisting of 5 female and 5 male native English speakers, and the number of combinations of source and target speakers was 90. The number of sentences uttered by each speaker was 216. The sampling frequency was set to 16 kHz.

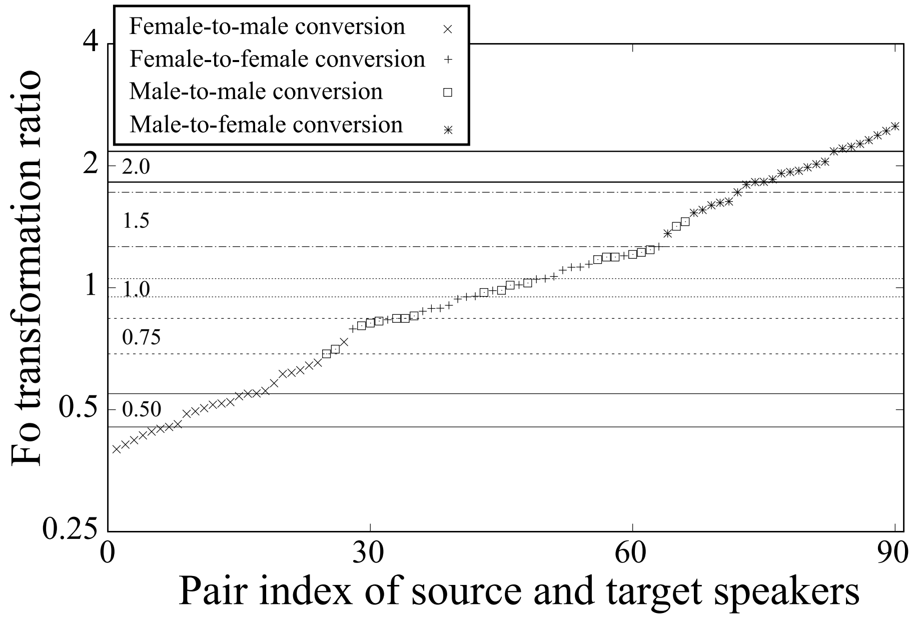


Figure 3.13.: F_0 transformation ratios between source and target speakers.

STRAIGHT [12] was used to extract spectral envelope, which was parameterized into the 1-24th mel-cepstral coefficients as the spectral feature. The frame shift was 5 ms. The mel log spectrum approximation (MLSA) filter [53] was used as the synthesis filter. As the source excitation features, we used F_0 and aperiodic components extracted with STRAIGHT [49]. The aperiodic components were averaged over five frequency bands, i.e., 0-1, 1-2, 2-4, 4-6, and 6-8 kHz, to be modeled with the GMM.

We investigated F_0 transformation ratios for all speaker possible pairs from 10 evaluation speakers (i.e., 45 speaker pairs in total) as shown in Figure 3.13, and selected 10 speaker pairs in each quantized F_0 transformation ratio (0.5, 0.75, 1.0, 1.5, and 2.0) as the source and target speaker pairs. We used 162 sentences for training and the remaining 54 sentences were used for evaluation. The speaker-dependent GMMs were separately trained for the individual source and target speaker pairs. We performed MS-based postfilter for the converted mel-cepstrum differential described in Section 3.2.2. The number of mixture components for the mel-cepstral coefficients was 128 and for the aperiodic components was 64. The number of subjects was 8 and they were not native English speakers.

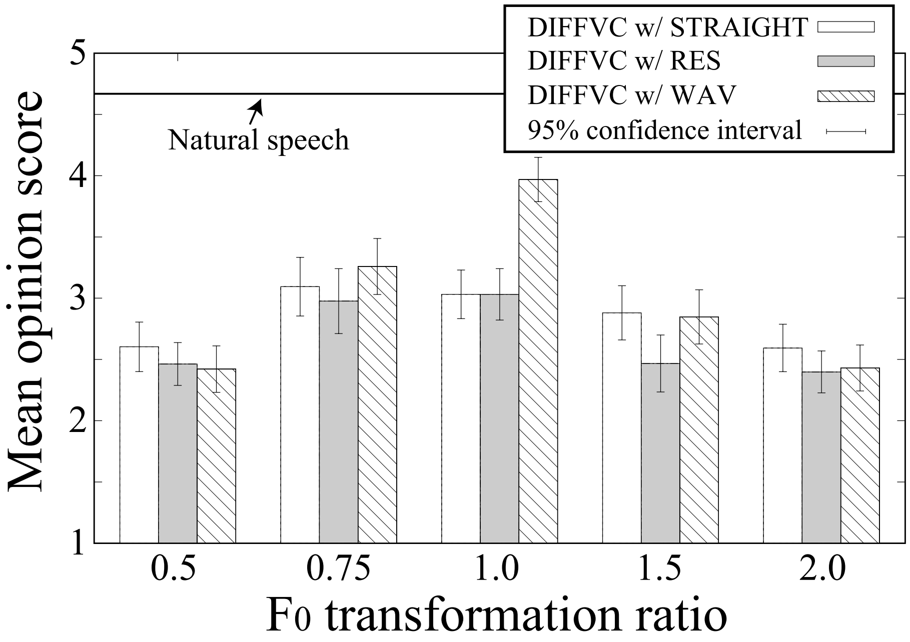


Figure 3.14.: Evaluation of sound quality of converted voice for intra/inter-gender DIFFVC.

Subjective evaluation

Two subjective evaluations were conducted. In the first test, we evaluated the sound quality of the converted voices using a mean opinion score (MOS). The natural and converted voice samples generated by three different DIFFVC methods were presented to subjects in random order. The subjects rated the quality of the converted voice using a 5–point scale: “5” for excellent, “4” for good, “3” for fair, “2” for poor, and “1” for bad. The number of evaluation sentences in each subject was 128.

In the second test, conversion accuracy in speaker identity was evaluated. In this test, F_0 transformation ratios were set to 0.5, 1.0, and 2.0. A natural voice sample of the target speaker was presented to the subjects first as a reference. Then, the converted voice samples generated by three different DIFFVC methods for the same sentences were presented in random order. The subjects selected which sample was more similar to the reference natural voice in terms of speaker identity. Each subject evaluated 90 sample pairs. They were allowed to replay each sample pair as many times as necessary.

Figure 3.14 indicates the results of the MOS test for sound quality. We can see a general tendency that sound quality degradation is caused by setting the F_0 transformation ratio to higher/lower values in all methods. When the F_0 transformation ratio is set to around 1.0, DIFFVC w/ WAV can achieve the highest sound quality. The sound quality achieved by DIFFVC w/ WAV rapidly degrades when setting the F_0 transformation ratio to higher or lower values than 1.0. On the other hand, DIFFVC w/ STRAIGHT and DIFFVC w/ RES tend to make such a quality degradation more gradually compared with DIFFVC w/ WAV. Nevertheless, the sound quality achieved by DIFFVC w/ WAV is still comparable to the other methods even if setting the F_0 transformation ratio to around 0.5 or 2.0. As for a comparison between DIFFVC w/ STRAIGHT and DIFFVC w/ RES, we can see that DIFFVC w/ STRAIGHT is slightly better than DIFFVC w/ RES when setting the F_0 transformation ratio to higher values (i.e., around 1.5 and 2.0). These results demonstrate that DIFFVC w/ WAV outperforms DIFFVC w/ STRAIGHT and DIFFVC w/ RES in terms of sound quality of the converted voices.

Figures 3.15 (a), (b) and (c) indicate the results of the preference test for speaker identity. We can see a tendency similar to that observed in the previous test on the converted sound quality; i.e., 1) DIFFVC w/ WAV yields better conversion accuracy for speaker identity than the other methods when setting the F_0 transformation ratio to around 1.0; 2) DIFFVC w/ WAV is still comparable to the other methods even when setting the F_0 transformation ratio to around 0.5 and 2.0; and 3) as for a comparison between DIFFVC w/ STRAIGHT and DIFFVC w/ RES, DIFFVC w/ STRAIGHT yields better conversion accuracy for speaker identity when setting the F_0 transformation ratio to around 0.5 and 2.0. Therefore, DIFFVC w/ WAV outperforms the other methods in terms of conversion accuracy for speaker identity as well.

These results suggest that DIFFVC w/ WAV is the best approach to implementing F_0 transformation to the DIFFVC framework in terms of both converted sound quality and conversion accuracy for speaker identity. Note that DIFFVC w/ WAV can also significantly reduce a computational cost in conversion.

3.5. Summary of the experimental evaluations

Figure 3.16 indicates a reference summery of the experimental results performed in Chapter 3. Although it is not possible to directly compare proposed techniques due

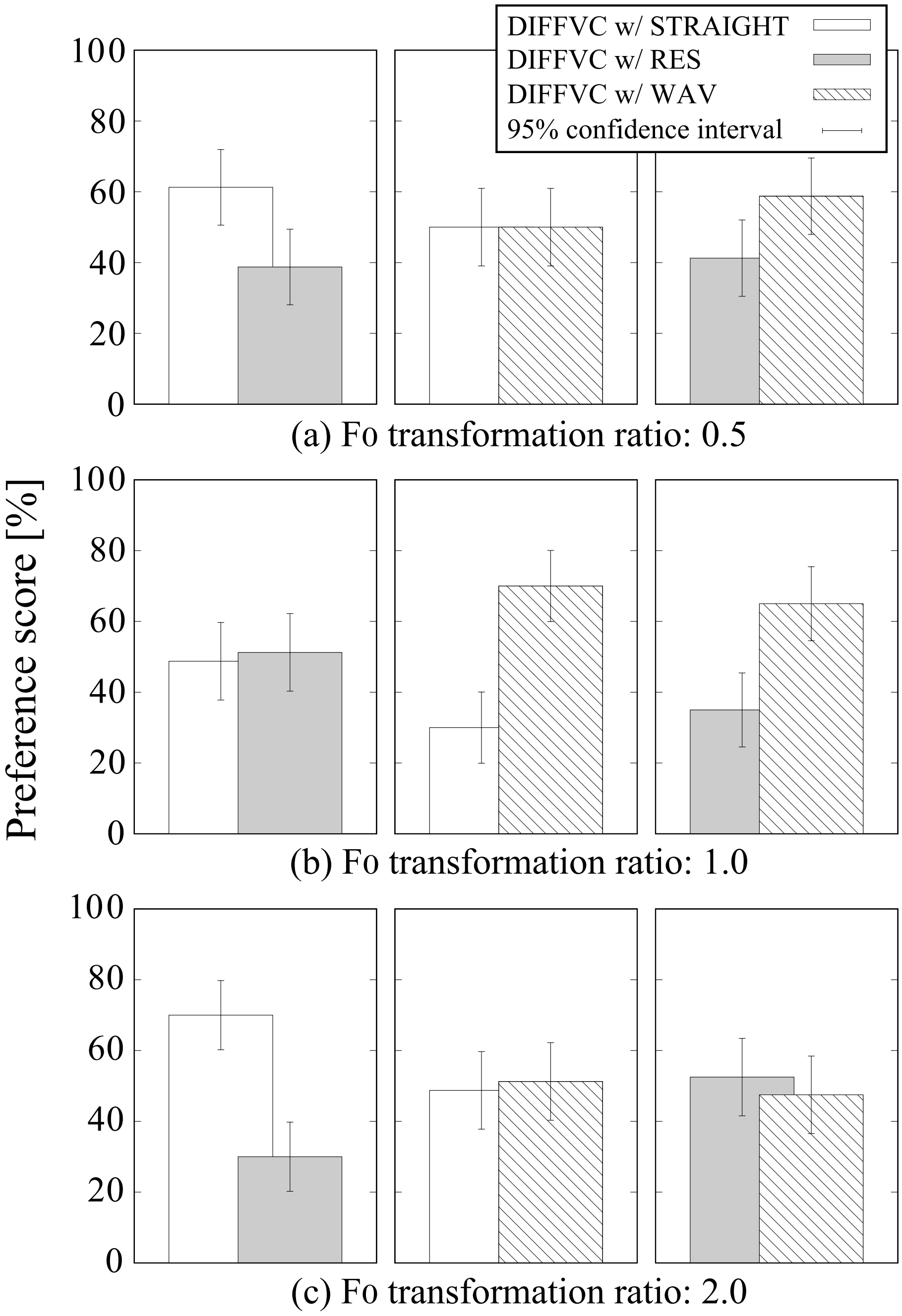


Figure 3.15.: Evaluation of conversion accuracy on speaker identity for intra/inter-gender DIFFVC.

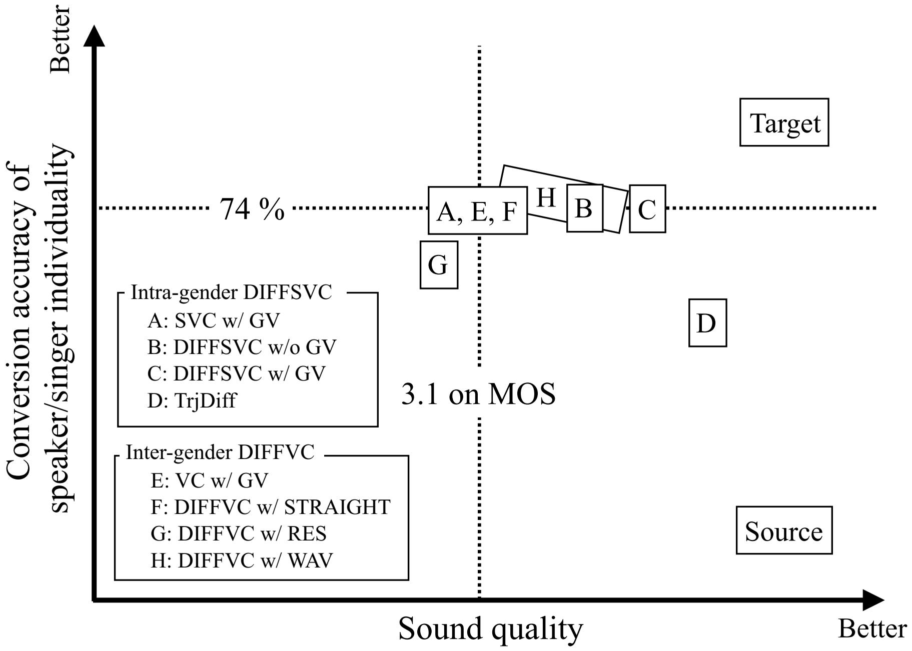


Figure 3.16.: A reference summery of the experimental evaluations in Chapter 3.

to various reasons such as different experimental conditions and evaluation methods, it is expected that this figure can help to understand the relationship between conventional and proposed techniques. In order to briefly compare the effectiveness of each proposed technique, we give a reference summery estimated from the experimental results described in Chapter 3 and Appendix A. In figure A.1 described in Appendix A, the NU-NAIST VC system achieved 3.1 opinion score and 74 % on conversion accuracy of the speaker individuality in Voice Conversion Challenge (VCC) 2016 where the NU-NAIST VC system almost equals to DIFFVC w/ STRAIGHT and its performance almost equals to conventional VC w/GV. Therefore, we can put a reference score to the figure. In the experimental results of proposed intra-gender DIFFSVC, DIFFSVC w/o GV and DIFFSVC w/ GV achieved higher sound quality and equivalent conversion accuracy compared with conventional SVC w/ GV. Although TrjDIff is possible to convert with higher sound quality, the conversion accuracy degrades significantly. In the experimental results of intra/inter-gender DIFFVC, when the F_0 transformation ratio is larger, DIFFVC w/ STRAIGHT and DIFFVC w/ WAV achieved better sound quality and conversion accuracy compared with DIFFVC w/ RES. Moreover, when the F_0 transformation ratio is lower, DIFFVC w/ WAV has higher sound quality and conversion accuracy compared with DIFFVC w/ STRAIGHT.

3.6. Summary

In this chapter, in order to improve the sound quality of the converted voice in statistical VC for speaking and singing voice, we have proposed a statistical waveform modification technique to convert voice timbre of a source speaker/singer into that of a target speaker/singer without using waveform generation based on vocoding.

Section 3.2: This section has described intra-gender statistical waveform modification technique based on spectral differential. At first, we have shown the technique to estimate the joint probability density function of the DIFFGMM using previously trained GMM. Then, in order to alleviate the sound quality degradation caused by the over-smoothing effect of the converted feature trajectory, we have proposed techniques to compensate the GV and MS of the converted feature trajectory. Moreover, we have proposed a differential parameter generation technique based on trajectory differential using intra/inter-speaker GMMs.

Section 3.3: In this section, to make it possible to apply intra-gender DIFFSVC framework into intra/inter-gender DIFFVC framework, we have proposed several F_0 transformation techniques. We proposed following F_0 transformation techniques: 1) F_0 transformation technique using the STRAIGHT vocoder, 2) F_0 transformation technique based on residual signala modification using WSOLA and resampling, and 3) F_0 transformation technique based on waveform modification using WSOLA and resampling.

Section 3.4: In this section, we have performed several experimental evaluations for intra-gender DIFFSVC and intra/inter-gender DIFFVC techniques. For the intra-gender DIFFSVC, the proposed techniques have achieved higher sound quality with equivalent conversion accuracy compared with the conventional SVC considering GV. For the intra/inter-gender DIFFVC, the proposed methods with or without using vocoding technique make it possible to apply intra-gender DIFFSVC framework into intra/inter-gender DIFFVC framework. The experimental results confirmed that the DIFFVC w/ WAV achieved higher sound quality for inter-gender conversion and equivalent sound quality for intra-gender conversion. In conclusion, the overall summary illustrates that statistical waveform modification techniques make it possible to convert higher sound quality and equivalent conversion accuracy on speaker/singer identity compared with those of the conversion VC based on GMM.

4. Voice timbre control via statistical waveform modification

4.1. Introduction

The singing voice is one of the most expressive components in music. In addition to pitch, dynamics, and rhythm, the linguistic information of the lyrics can be used by singers to express more varieties of expression than other music instruments. Although singers can also expressively control their voice characteristics such as voice timbre to some degree, they usually have difficulty in changing their own voice characteristics widely, (e.g. changing them into those of another singer's singing voice) owing to physical constraints in speech production. If it would be possible for singers to freely control voice characteristics beyond these physical constraints, it will open up entirely new ways for singers to express themselves.

Singing synthesis system [54–58] has been a growing interest in computer-based music technology to generate an arbitrary singing voice. Entering notes and lyrics to the singing synthesis engine, users (e.g., composers) can easily produce a synthesized singing voice which has a specific singer's voice characteristics, different from those of the users. Previous work has proposed techniques to flexibly control the synthesized singing voice as the users want by automatically [59, 60] adjusting parameters of the singing synthesis system to generate more expressive synthesized singing voice. Although these technologies are effective to produce the singing voices designed by the users, it is essentially difficult to produce synthesized singing voices by controlling all singing voice components including lyrics on the fly.

In previous research, a number of techniques have been proposed to change the characteristics of singing voices. One typical method is singing voice conversion based on speech morphing in the speech analysis/synthesis framework [61]. This method makes it possible to independently morph several acoustic parameters, such as spectral envelope, F_0, and duration, between singing voices of different singers or different singing styles. One of the limitations of this method is that the morphing can only be applied to singing voice samples of the same song.

To make it possible to more flexibly change singing voice characteristics, a singing VC (SVC) technique enable us to convert the source singer's singing voice into another target singer's singing voice [9, 62]. Moreover, SVC based on eigenvoice Gaussian mixture model (EV-GMM) have been proposed as a technique to convert an arbitrary source singer into an arbitrary target signer [15, 63]. Although this technique is also capable of flexibly changing singing voice timbre by manipulating the adaptation parameters even if no target singing voice sample is available, it is difficult to find the ideal singing voice timbre, because it is hard to predict the change of singing characteristics caused by the manipulation of each adaptation parameter. In the area of statistical parametric speech synthesis [38], there have been several attempts at developing techniques for manually controlling voice characteristics of synthetic speech by manipulating intuitively controllable parameters [64–66]. A similar method has also been proposed in statistical VC [16] with multiple-regression GMM (MR-GMM). Although these methods have only been applied to voice characteristics control for normal speech, it is expected that they would also be effective for controlling singing voice characteristics.

In order to implement intuitive voice timbre control framework using perceptually understandable cues, we focus on the perceived age or the age that a listener predicts the singer to be, of singing voices as one of the factors to intuitively describe the singing voice. However, the perceived age of singing voices has not yet been studied deeply. As fully understanding the acoustic features that contribute to the perceived age of singing voices is essential to the development of VC techniques to modify a singer's perceived age,

In this chapter, we first perform an investigation of the acoustic features that play a part in the listener's perception of the singer's age at first. We conduct several types of perceptual evaluation to investigate 1) how well the perceived age of singing voices corresponds to the actual age of the singer, 2) whether or not singing VC processing causes adverse effects on the perceived age of singing voices, 3) which spectral or prosodic features have a larger effect on the perceived age, and 4) which spectral or

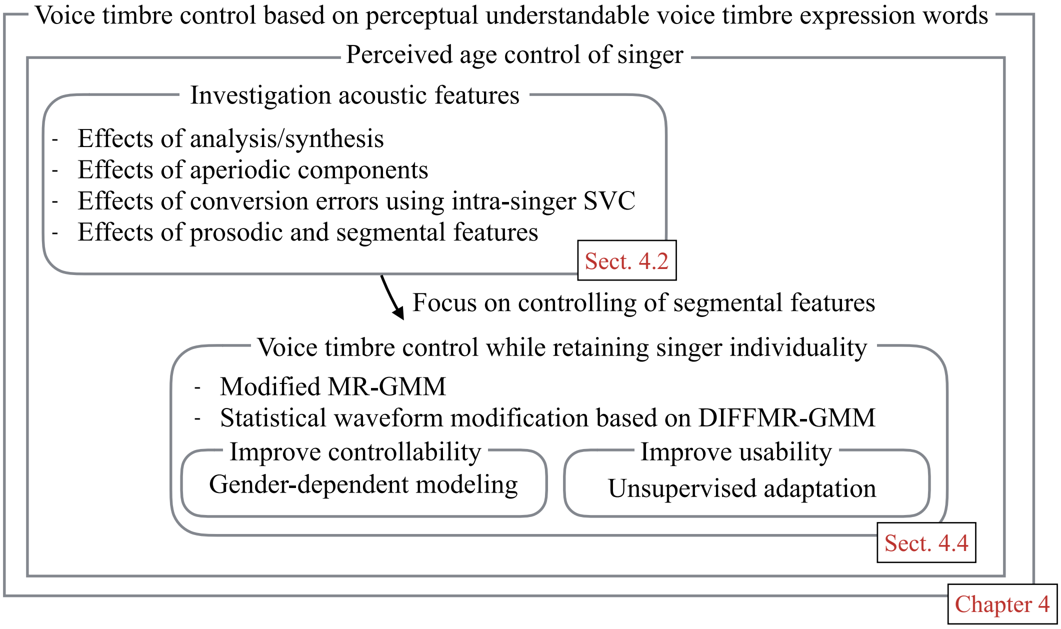

Voice timbre control based on perceptual understandable voice timbre expression words
Perceived age control of singer
Investigation acoustic features
- Effects of analysis/synthesis
- Effects of aperiodic components
- Effects of conversion errors using intra-singer SVC
- Effects of prosodic and segmental features
Sect. 4.2
Focus on controlling of segmental features
Voice timbre control while retaining singer individuality
- Modified MR-GMM
- Statistical waveform modification based on DIFFMR-GMM
Improve controllability
Gender-dependent modeling
Improve usability
Unsupervised adaptation
Sect. 4.4
Chapter 4

Figure 4.1.: The rest of Chapter 4.

prosodic features have an individuality of a singer. Then, we propose a novel voice timbre conversion method that converts the singer's perceived age while maintaining individuality in SVC. We propose a technique to control the perceived age while retaining singer individuality. Moreover, towards the development of a better controllable, higher-quality, and more flexible framework, we also propose the following three methods for the perceived age control technique; 1) a method using gender-dependent MR-GMMs, 2) a method using direct waveform modification based on spectrum differential, and 3) a rapid unsupervised adaptation method.

The rest of this chapter is shown in Figure 4.1. In Section 4.2, we investigate acoustic features affecting on the perceived age of the singer. We investigate following the effects 1) effects of analysis/synthesis, 2) effects of aperiodic components, 3) effects of aperiodic components, 4) effects of conversion errors using intra-singer SVC, and 5) effects of prosodic and segmental features. In Section 4.4, we propose several techniques to control voice timbre based on the perceived age while retaining singer individuality. At fast, in order to control voice timbre while retaining the singer identity, we propose a voice timbre control technique based on the Modified MR-GMM.

Then, in order to improve perceived age controllability and sound quality, we propose a voice timbre control technique based on a gender-dependent Modified differential MR-GMM (DIFFMR-GMM). Finally, we propose an unsupervised adaptation technique to easily develop the Modified DIFFMR-GMM for an arbitrary source signer. In Section 4.5, we evaluate the effectiveness of the voice timbre control techniques proposed in this chapter. In Section 4.6, we summarize this chapter.

4.2. Investigation of the acoustic features affecting perceived age

In the traditional SVC [9, 15], only the spectral features such as mel-cepstrum are converted. It is also straightforward to convert the aperiodic components (ACs) [49], which capture noise strength on each frequency band of the excitation signal, as in the traditional VC for natural voices [13]. If the perceived age of singing voices is captured well by these acoustic features, it will make it possible to develop a real-time SVC system capable of controlling the perceived age of singing voices by combining SVC with MR-GMM (described in Section 4.4.1) and real-time statistical VC techniques [24,25]. On the other hand, if the perceived age of singing voices is not captured at all by these acoustic features, which mainly represent segmental features, the conversion of other acoustic features, such as prosodic features (e.g., F_0 pattern), will also be necessary. In such a case, the voice characteristics control framework of HMM-based speech synthesis [64, 66] can be used in the SVC system to control the perceived age of singing voices, although it is not straightforward to develop a real-time SVC system in this framework. In this section, we compare the perceived age of natural singing voices with that of several types of synthesized singing voices by modifying acoustic features as shown in Table 4.1 for the purpose of investigating acoustic features affecting the perceived age in singing voices to clarify which types of techniques can be implemented for the SVC system.

4.2.1. Effects of analysis/synthesis

In the analysis/synthesis framework, a voice is first converted into parameters of a source-filter model, then simply re-synthesized into a waveform using these parameters

Table 4.1.: Acoustic features of several types of synthesized singing voices.

Features	Power, F_0, duration	Mel-cepstrum	Aperiodic components
Analysis/synthesis (w/ ACs)	Source singer	Source singer	Source singer
Analysis/synthesis (w/o ACs)	Source singer	Source singer	Removed
Intra-singer SVC (source)	Source singer	Converted to source singer	Converted to source singer
Intra-singer SVC (target)	Target singer	Converted to target singer	Converted to target singer
SVC	Source singer	Converted to target singer	Converted to target singer

without change. We define this re-synthesized singing voice as analysis/synthesis (w/ ACs). As analysis and synthesis are necessary steps in converting acoustic features of singing voices, we investigate the effects of distortion caused by analysis/synthesis on the perceived age of singing voices. We use STRAIGHT [12] as a widely used high-quality analysis/synthesis method to extract acoustic features consisting of the spectral envelope, F_0, and ACs. The spectral envelope is further parameterized with mel-cepstrum.

4.2.2. Effects of aperiodic components

As mentioned above, previous research [31] has shown that ACs tend to change with aging in normal speech as mentioned above. We investigate the effects of ACs on the perceived age of singing voices. Analysis/synthesized singing voice samples are reconstructed from mel-cepstrum and F_0 extracted with STRAIGHT. In synthesis, only a pulse train with phase manipulation [12] instead of STRAIGHT mixed excitation [13] is used to generate voiced excitation signals. We define this re-synthesized singing voice as analysis/synthesis (w/o ACs).

4.2.3. Effects of conversion errors

In SVC, conversion errors are inevitable. For example, some detailed structures of acoustic features not well modeled by the GMM of the joint probability density and often disappear through the statistical conversion process. Therefore, the acoustic space on which the converted acoustic features are distributed tends to be smaller than the acoustic space that of the natural acoustic features. We investigate the effect of the conversion errors caused by this acoustic space reduction on the perceived age of singing voices by converting one singer's singing voice into the same singer's singing voice.

This SVC process is called intra-singer SVC (source/target).

To achieve intra-singer SVC (source/target) for a specific singer, we must create a GMM to model the joint probability density of the same singer's acoustic features, i.e., $P(\boldsymbol{X}_t, \boldsymbol{X}'_t|\boldsymbol{\lambda})$ where $\boldsymbol{X}_t$ and $\boldsymbol{X}'_t$ respectively show the source and target acoustic features of the same singer. It is impossible to train such a GMM by simply using the source feature vector of the source singer $\boldsymbol{X}_t$ as the target feature vector $\boldsymbol{Y}_t$ because this duplication causes the rank deficiency of the covariance matrix. Namely, the following conditions need to hold; $\boldsymbol{X}_t$ is different from $\boldsymbol{X}'_t$; they depend on each other, and both are identically distributed. This GMM can be trained using a parallel data set consisting of the song pairs sung by the source singer but the source singer needs to sing the same songs twice to develop such a parallel data set. As a more convenient way to develop the GMM for intra-singer SVC (source/target), we use the framework of many-to-many EVC. The GMM is analytically derived from the GMM of the joint probability density of the acoustic features of the same singer and another reference singer, i.e., $P(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda})$ where $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ respectively show the source feature vector of the same singer and that of the reference singer, by marginalizing out the acoustic features of the reference singer in the same manner as used in the many-to-many EVC as follows:

$$\begin{aligned} P\left(\boldsymbol{X}_t, \boldsymbol{X}'_t|\boldsymbol{\lambda}\right) &= \sum_{m=1}^{M} P\left(m|\boldsymbol{\lambda}\right) \int P\left(\boldsymbol{X}_t|\boldsymbol{Y}_t, m, \boldsymbol{\lambda}\right) \\ &\qquad P\left(\boldsymbol{X}'_t|\boldsymbol{Y}_t, m, \boldsymbol{\lambda}\right) P\left(\boldsymbol{Y}_t|m, \boldsymbol{\lambda}\right) \mathrm{d}\boldsymbol{Y}_t \\ &= \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix} \boldsymbol{X}_t \\ \boldsymbol{X}'_t \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(X)} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XYX)} \\ \boldsymbol{\Sigma}_m^{(XYX)} & \boldsymbol{\Sigma}_m^{(XX)} \end{bmatrix} \right), \end{aligned} \tag{4.1}$$

$$\boldsymbol{\Sigma}_m^{(XYX)} = \boldsymbol{\Sigma}_m^{(XY)} \boldsymbol{\Sigma}_m^{(YY)^{-1}} \boldsymbol{\Sigma}_m^{(YX)}. \tag{4.2}$$

Using this GMM, intra-singer SVC (source/target) is performed in the same manner as described in Section 2.3.2. The converted singing voice sample essentially has the same singing voice characteristics as those before the conversion although they suffer from conversion errors. We define this converted singing voice as intra-singer SVC (source/target).

4.2.4. Effects of prosodic and segmental features

To investigate which acoustic features have a larger effect on the perceived age of singing voices, segmental features or prosodic features, we use the SVC for converting only segmental features, such as mel-cepstrum and ACs, of a source singer into those of a different target singer. The converted singing voice samples essentially have the segmental features of the target singer and the prosodic features, such as F_0 patterns, power patterns, and duration, of the source singer.

4.3. Experimental evaluation for investigation

4.3.1. Experimental condition

In our experiments, we first investigated the correspondence between the perceived age and the actual age of the singer. We used the AIST humming database [67] consisting of singing voices of 25 songs with Japanese lyrics sung by Japanese male and female amateur singers in their 20s, 30s, 40s, and 50s. The total number of singers in the database was 75. The length of each song was approximately 20 seconds. For evaluation, only one Japanese song (No. 39) was used. Eight Japanese male subjects in their 20s were asked to guess the age of each singer by listening to his/her singing voices.

In the second experiment, we investigated the acoustic features that affect the perceived age of singing voices. We did so by comparing the perceived age of natural singing voices with that of each type of synthesized singing voice as shown in Table 4.1. Eight Japanese male subjects in his 20s assigned the perceived age to each synthesized singing voice. We selected 16 singers consisting of four singers (two male singers and two female singers) from each age group, i.e., their 20s, 30s, 40s, or 50s as evaluation singers. The singers were also separated into two groups, A and B, so that one group always included one male singer and one female singer in each age group. The subjects in each group evaluated only singing voices of the corresponding singer group.

In the third experiment, we investigated which acoustic features more affected the singer's individuality of singing voices. We divided the 16 evaluation singers into four groups, M1, M2, F1 and F2, so that each group included four male or female singers from all age groups. The subjects were also randomly separated into four groups. Converted singing voices with SVC were created in every combination of source and target singer pairs in each group (i.e., 12 combinations) as evaluation samples. Converted singing voices with intra-singer SVC (source/target) were also created for individual singers (four male or female singers) in each group as reference samples. The subjects were asked to separate the evaluation samples into four classes in accordance with the reference samples on the basis of similarity of singer's individuality. The subjects were allowed to listen to the evaluation and reference samples as many times as they wanted. We gave instructions to the subjects to evaluate the singer's individuality considering a possibility of changes of singing voice characteristics caused by aging.

The sampling frequency was set to 16 kHz. The 1st through 24th mel-cepstral coefficients extracted by STRAIGHT analysis were used as spectral features. As the source excitation features, we used F_0 and ACs in five frequency bands, i.e., 0–1, 1–2, 2–4, 4–6, and 6–8 kHz, which were also extracted by STRAIGHT analysis. The frame shift was 5 ms.

As training data for the GMMs used in intra-singer SVC (source/target) and SVC, we used 18 songs including the evaluation song (No. 39). In the intra-singer SVC (source/target), GMMs for converting the mel-cepstrum and ACs were trained for each of the selected 16 singers. Another singer not included in these 16 singers was used as the reference singer to create each parallel data set for the GMM training. In the SVC, the GMMs for converting mel-cepstrum and ACs were trained for all combinations of the source and target singer pairs in each singer group. The numbers of mixture components of each GMM were optimized experimentally.

4.3.2. Experimental result

Comparison between perceived age and actual age

Figure 4.2 indicates the correlation between the perceived age of natural singing voices and the actual age of the singer. Each point indicates the perceived age of each singer averaged over all subjects. The standard deviation of the perceived age in each singer over all subjects is 6.17. The correlation coefficient between the perceived age and the actual age in this figure is 0.81. These results show a quite high correlation between the perceived age and the actual age.

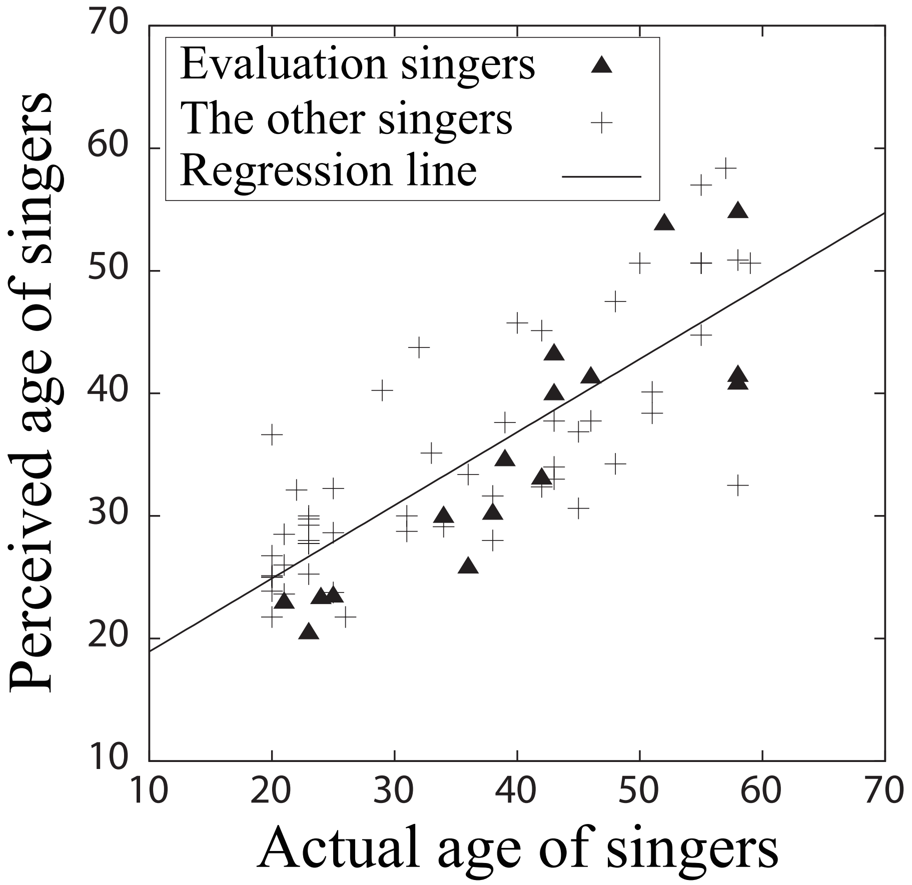


Figure 4.2.: Correlation between singer's actual age and perceived age.

Table 4.2.: Differences of the perceived age between natural singing voices and each type of the synthesized singing voices.

Methods	Average	Standard deviation	Correlation coefficient
Analysis/synthesis (w/ ACs)	0.77	3.57	0.96
Analysis/synthesis (w/o ACs)	0.44	3.58	0.96
Intra-singer SVC	-0.50	7.25	0.85

Acoustic features affecting perceived age

Table 4.2 indicates average values and standard deviations of differences between the perceived age of natural singing voices and each type of intra-singer synthesized singing voice: analysis/synthesis (w/ ACs), analysis/synthesis (w/o ACs) and the intra-singer SVC (source/target). The table also indicates correlation coefficients between the perceived age of natural and synthesized voices. From the results, we can see that in analysis/synthesis (w/ ACs), the perceived age difference is small and the correlation coefficient is very high. Therefore, distortion caused by analysis/synthesis processing does not affect the perceived age. It can be observed from analysis/synthesis (w/o ACs) that this result does not change even if not using ACs. Therefore, ACs do not affect the perceived age of singing voices. On the other hand, intra-singer SVC (source/target) causes slightly larger differences between natural singing voices and the synthesized singing voices. Therefore, some acoustic cues to the perceived age are removed through the statistical conversion processing. Nevertheless, the perceived age differences are relatively small, and therefore, it is likely that important acoustic cues to the perceived age are still kept in the converted acoustic features.

Figures 4.3 and 4.4 indicate a comparison between the perceived age of singing voices generated by SVC and intra-singer SVC (source/target). In each figure, the vertical axis indicates the perceived age of converted singing voices by SVC (prosodic features: source singer, segmental features: target singer). The horizontal axis in Figure 4.3 indicates the perceived age of singing voices generated by intra-singer SVC (source) and that in Figure 4.4 indicates the perceived age of singing voices generated by intra-singer SVC (target). Therefore, if the prosodic features more strongly affect the perceived age than the segmental features, a higher correlation will be observed in Figure 4.3. If the segmental features more strongly affect the perceived age than the prosodic features, a higher correlation will be observed in Figure 4.4 than in Figure 4.3. These figures demonstrate that 1) the segmental features affect the perceived age but the effects are limited as shown in positive but weak correlation in Figure 4.4 and 2) the prosodic features have a larger effect on the perceived age than the segmental features.

Acoustic features affecting singer individuality

In this experiment, we investigated which prosodic and segmental features have a larger impact on singer's individuality. Table 4.3 indicates the ratios judged by subjects based on similarity of between the converted singing voice from the source singer into the target singer with SVC and the source, target, or other singers' reference singing voices that were generated by intra-singer SVC (source/target). If the prosodic features

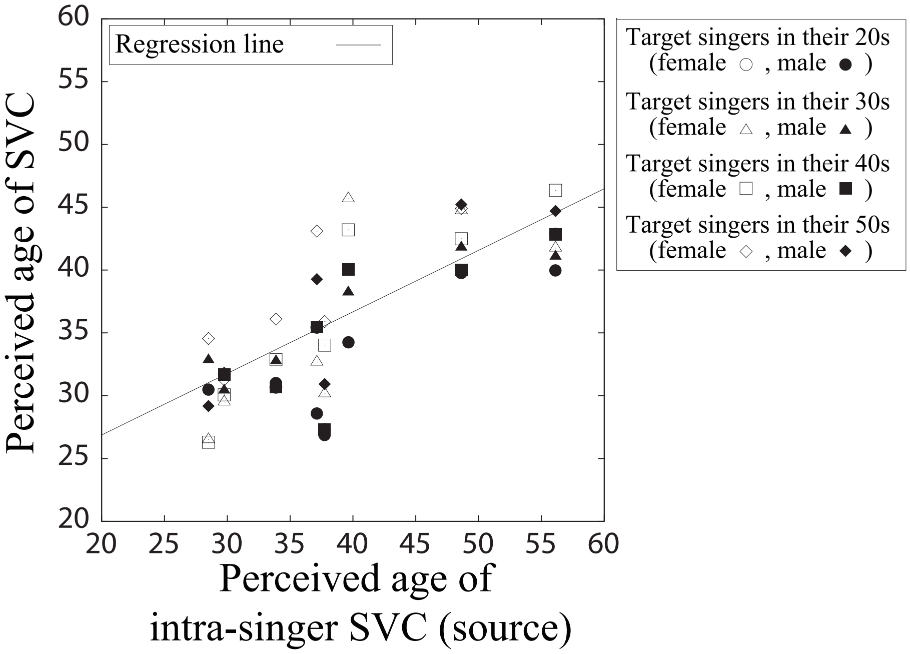


Figure 4.3.: Correlation of perceived age between singing voices generated by the intra-singer SVC (source) and the SVC.

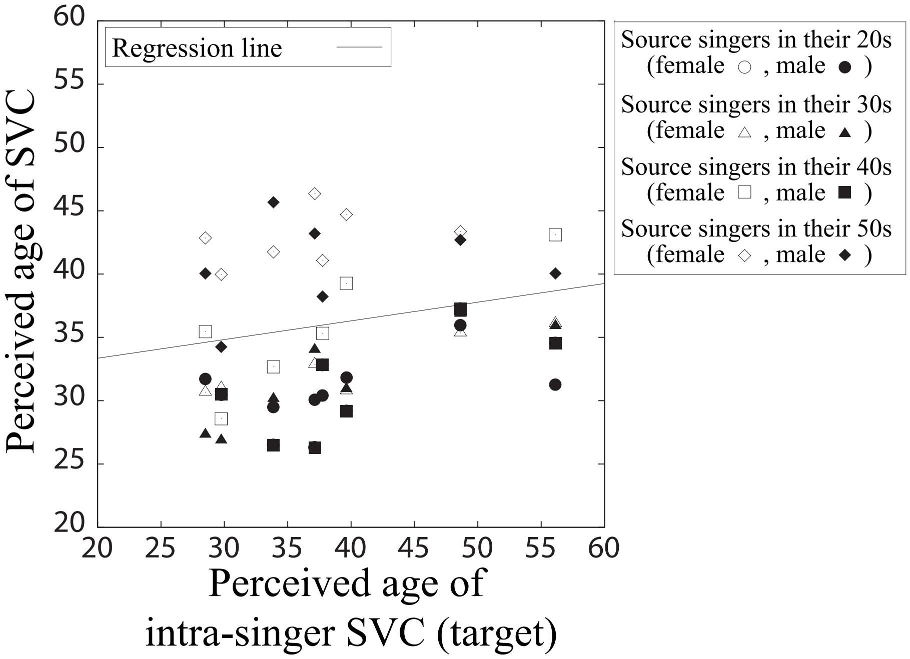


Figure 4.4.: Correlation of perceived age between singing voices generated by the intra-singer SVC (target) and the SVC.

Table 4.3.: Evaluation of singer identification in SVC.

Acoustic features	Ratio
Prosodic features	52.08
Segmental features	35.42
Disagreement	12.50

more strongly have the individuality of singer than segmental features, then singing voice converted with SVC is classified into intra-singer SVC (source). On the other hand, if the segmental features more strongly have the individuality of singer than prosodic features, then the singing voice converted with SVC is classified into intra-singer SVC (target). If the singing voice converted with SVC is classified to the other singers' reference singing voices, it was counted as a disagreement sample. This table demonstrates that individuality of a singer is distinguished from prosodic features rather than segmental features. This result has a similar tendency on Figures 4.3 and 4.4. Namely, there is a correlation between singer's individuality and perceived age. These results suggest that if it is necessary to make large changes in the perceived age, then prosodic features are the most suitable acoustic features. However, it will also cause changes of singer's individuality. In contrast, if it is required to change only the perceived age while remaining singer's individuality, segmental features are more appropriate features although a range of changes of the perceived age is limited.

4.4. Voice timbre control based on perceived age while retaining singer identity

In the last section, we indicated that segmental features are suitable to control the perceived age to retain singer individuality. In this section, we develop a perceived age controllable SVC technique for a specific singer. At first, VC based on the MR-GMM is applied to SVC to convert segmental features by manipulating the perceived age. Moreover, we propose a modified MR-GMM to maintain the singer's individuality.

To improve controllability of the perceived age, sound quality of the converted voice, and flexibility of the model development in the conventional voice timbre control method, we further implement three techniques, 1) gender-dependent MR-GMMs for more accurately capturing spectral variations depending on the perceived age, 2) direct waveform modification based on spectral differential, and 3) a rapid unsupervised adaptation method based on maximum a posteriori (MAP) estimation to easily develop the singer-dependent MR-GMM.

4.4.1. Modified MR-GMM implementation based on many-to-many SVC

SVC with MR-GMM also consists of a training process and a conversion process. The MR-GMM is trained using multiple parallel data sets consisting of the source singer's singing voices and many pre-stored target singers' singing voices. The joint probability density of $2D$-dimensional joint static and dynamic feature vectors modeled by the MR-GMM is given by

$$\begin{aligned} & P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)} | \boldsymbol{\lambda}^{(MR)}, w^{(s)}\right) \\ & = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix} \boldsymbol{X}_t \\ \boldsymbol{Y}_t^{(s)} \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)}(s) \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix}\right), \end{aligned} \tag{4.3}$$

where $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$ and $\boldsymbol{Y}_t^{(s)} = [\boldsymbol{Y}_t^{(s)\top}, \Delta\boldsymbol{Y}_t^{(s)\top}]^\top$ show static and delta feature vectors of the source and s-th pre-stored target singer. The mean vector of the s-th pre-stored target singer is given by

$$\boldsymbol{\mu}_m^{(Y)}(s) = \boldsymbol{b}_m^{(Y)} w^{(s)} + \overline{\boldsymbol{\mu}}_m^{(Y)}, \tag{4.4}$$

where $\boldsymbol{b}_m^{(Y)}$ and $\overline{\boldsymbol{\mu}}_m^{(Y)}$ indicate the representative vector and bias vector respectively. $w^{(s)}$ indicates the s-th pre-stored target singer's perceived age score, which is manually assigned for each pre-stored target singer.

In the conversion process, the perceived age score is manually set to the desired value. Then, the converted feature vector is determined in the same manner as described in 2.3.2.

To make it easier to develop the MR-GMMs for individual source singers (i.e., users), we apply the framework of many-to-many SVC [15] to SVC based on MR-

GMM. The joint probability density of many-to-many MR-GMM follows:

$$
\begin{aligned}
& P\left(\boldsymbol{Y}_t^{(i)}, \boldsymbol{Y}_t^{(o)} | \boldsymbol{\lambda}^{(MR)}, w^{(i)}, w^{(o)}\right) \\
& = \sum_{m=1}^{M} P\left(m|\boldsymbol{\lambda}^{(MR)}\right) \int P\left(\boldsymbol{Y}_t^{(i)} | \boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(MR)}, w^{(i)}\right) \\
& \quad P\left(\boldsymbol{Y}_t^{(o)} | \boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(MR)}, w^{(o)}\right) P\left(\boldsymbol{X}_t | m, \boldsymbol{\lambda}^{(MR)}\right) \mathrm{d}\boldsymbol{X}_t \\
= & \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix} \boldsymbol{Y}_t^{(i)} \\ \boldsymbol{Y}_t^{(o)} \end{bmatrix} ; \begin{bmatrix} \boldsymbol{\mu}_m^{(Y)}(i) \\ \boldsymbol{\mu}_m^{(Y)}(o) \end{bmatrix} , \begin{bmatrix} \boldsymbol{\Sigma}_m^{(YY)} & \boldsymbol{\Sigma}_m^{(YXY)} \\ \boldsymbol{\Sigma}_m^{(YXY)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix} \right),
\end{aligned}
\tag{4.5}
$$

$$
\boldsymbol{\Sigma}_m^{(YXY)} = \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(XX)-1} \boldsymbol{\Sigma}_m^{(XY)}, \tag{4.6}
$$

where $w^{(i)}$ and $w^{(o)}$ indicate the perceived age score of the source singer and that of the target singers, respectively. Source and target mean vectors are given by Eq. (4.4).

It is possible to use Eq. (4.4) to describe the input mean vectors $\boldsymbol{\mu}_m^{(Y)}(i)$ based on the perceived age score of the source singer. However, an ccuracy of acoustic modeling by the MR-GMM tends to decrease since the acoustic characteristics of the source singer are not always modeled well on a subspace spanned by the basis vector. To develop a better MR-GMM for the source singer, we assume an ideal condition that singing voice samples of the source singer in accordance with those of the reference singer that are used in the MR-GMM training are available. Namely, we suppose that it is possible to prepare a parallel data set of each user and the reference singer. This condition is still practical in the development of the user-dependent SVC system. Using a parallel data of the source singer's singing voice and the reference singer's singing voice, the input mean vector of the MR-GMM is updated in the sense of a maximum likelihood criterion. Consequently, the input mean vector is given by

$$
\boldsymbol{\mu}_m^{(Y)}(i) = \hat{\boldsymbol{\mu}}_m^{(Y)}, \tag{4.7}
$$

where $\hat{\boldsymbol{\mu}}_m^{(Y)}$ is its maximum likelihood estimate. Note that it is also possible to train all parameters of the MR-GMM using the parallel data sets of the user and all pre-stored target singers without using the many-to-many SVC framework. However, the training method presented here is still useful to effectively reduce computational cost to develop the MR-GMM because it is necessary to update only input mean vectors as shown in

Eq. (4.7). Moreover, there is a possibility to reduce the amount of singing voice data of the user used for training or implement an unsupervised training approach without the parallel data set based on model adaptation techniques.

In SVC with many-to-many MR-GMM, it is possible to convert voice timbre of the source singer into desired voice timbre in accordance with an output perceived age score. However, the output mean vector given by Eq. (4.4) only expresses average voice characteristics of several pre-stored target singers. Therefore, a converted singing voice doesn't express voice timbre of the source singer.

For the purpose of developing SVC based on perceived age while retaining the source singer's individuality, we change the representative form of the output mean vector as follows:

$$\begin{aligned}
\boldsymbol{\mu}_m^{(Y)}(o) &= \boldsymbol{b}_m^{(Y)} w^{(o)} + \overline{\boldsymbol{\mu}}_m^{(Y)} \\
&= \boldsymbol{b}_m^{(Y)} (w^{(i)} + \Delta w) + \overline{\boldsymbol{\mu}}_m^{(Y)} \\
&= \boldsymbol{b}_m^{(Y)} w^{(i)} + \overline{\boldsymbol{\mu}}_m^{(Y)} + \boldsymbol{b}_m^{(Y)} \Delta w \\
&\simeq \hat{\boldsymbol{\mu}}_m^{(Y)} + \boldsymbol{b}_m^{(Y)} \Delta w \qquad (4.8)
\end{aligned}$$

where the perceived age score of the output singing voice $w^{(o)}$ is represented by that of the input singing voice $w^{(i)}$ and a difference perceived age score Δw between them. In the modified representative form, the output mean vector is represented by the input mean vector $\hat{\boldsymbol{\mu}}_m^{(Y)}$ and the additional vector in accordance with a difference perceived age score Δw. As the input mean vector $\hat{\boldsymbol{\mu}}_m^{(Y)}$ is directly used instead of its projection on the subspace $\boldsymbol{b}_m^{(Y)} w^{(i)} + \overline{\boldsymbol{\mu}}_m^{(Y)}$, it is expected that acoustic characteristics of the source singer's singing voice are well preserved in this modified representative form.

4.4.2. Perceived age control via statistical waveform modification

As an SVC framework without using vocoder-based waveform generation, we have proposed a statistical waveform modification technique based on direct waveform modification using spectral differential in Chapter 3. In this section, this method is applied to the voice timbre control framework using the MR-GMM.

Figure 4.5 (b) shows proposed conversion processes based on Modified DIFFMR-GMM. In the direct waveform modification based on spectral differential, the spectral feature differential between the source singing voice and the converted singing voice

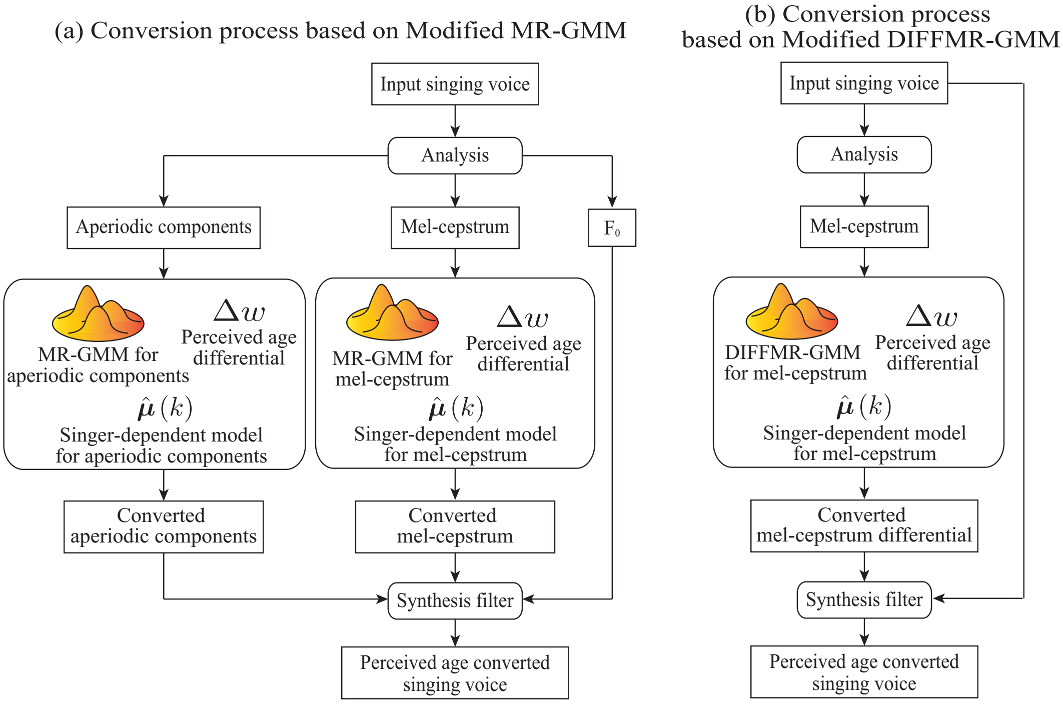

(a) Conversion process based on Modified MR-GMM
Input singing voice
Analysis
Aperiodic components
Mel-cepstrum
F₀
Δw
Perceived age differential
MR-GMM for aperiodic components
$\hat{\boldsymbol{\mu}}(k)$
Singer-dependent model for aperiodic components
MR-GMM for mel-cepstrum
Singer-dependent model for mel-cepstrum
Converted aperiodic components
Converted mel-cepstrum
Synthesis filter
Perceived age converted singing voice
(b) Conversion process based on Modified DIFFMR-GMM
DIFFMR-GMM for mel-cepstrum
Converted mel-cepstrum differential

Figure 4.5.: Conversion processes of perceived age control based on Modified MR-GMM and Modified DIFFMR-GMM.

is directly estimated from the source singer's spectral features using the DIFFMR-GMM (DIFFMR-GMM) modeling the joint probability density function of the source singer's spectral features and the spectral feature differential caused by the given perceived age differential. This differential model can be analytically derived from the conventional singer-dependent MR-GMM by applying a simple linear transform to the conventional model. The source singer's spectral feature is converted into the spectral feature differential using the DIFFMR-GMM.

Then, a waveform of the source singing voice is directly filtered with a time sequence of the estimated the spectral feature differentials. In this conversion process, the converted singing voice is free from various errors usually observed in the conventional waveform generation process with vocoder, such as F_0 extraction errors, unvoiced/voiced decision errors, spectral parameterization errors caused by liftering on the mel-cepstrum, and so on.

The DIFFMR-GMM is analytically derived from the singer-dependent MR-GMM as follows. Let $\boldsymbol{D}_t = \left[\boldsymbol{d}_t^\top, \Delta\boldsymbol{d}_t^\top\right]^\top$ denote the joint static and delta differential feature vector, where $\boldsymbol{d}_t = \boldsymbol{y}_t(o) - \boldsymbol{y}_t(i)$. The 2D-dimensional joint static and delta feature vector between the source and the differential features is represented as a linear transformation of the original joint feature vectors as follows:

$$\begin{bmatrix} \boldsymbol{Y}_t(i) \\ \boldsymbol{D}_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y}_t(i) \\ \boldsymbol{Y}_t(o) - \boldsymbol{Y}_t(i) \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ -\boldsymbol{I} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{Y}_t(i) \\ \boldsymbol{Y}_t(o) \end{bmatrix}, \tag{4.9}$$

where $\boldsymbol{I}$ denotes the identity matrix. Applying this linear transform to the singer-dependent MR-GMM, the DIFFMR-GMM is derived as follows:

$$\begin{aligned} & P\left(\boldsymbol{Y}_t(i), \boldsymbol{D}_t | \boldsymbol{\lambda}^{(DIFFMR)}, \hat{\boldsymbol{\mu}}(Y), \Delta w\right) \\ & = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix} \boldsymbol{Y}_t(i) \\ \boldsymbol{D}_t \end{bmatrix}; \begin{bmatrix} \hat{\boldsymbol{\mu}}_m^{(Y)} \\ \boldsymbol{b}_m^{(Y)} \Delta w \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_m^{(YY)} \boldsymbol{\Sigma}_m^{(DYD)} \\ \boldsymbol{\Sigma}_m^{(DYD)} \boldsymbol{\Sigma}_m^{(DD)} \end{bmatrix}\right), \end{aligned} \tag{4.10}$$

$$\boldsymbol{\Sigma}_m^{(DYD)} = \boldsymbol{\Sigma}_m^{(YXY)} - \boldsymbol{\Sigma}_m^{(YY)}, \tag{4.11}$$

$$\boldsymbol{\Sigma}_m^{(DD)} = 2(\boldsymbol{\Sigma}_m^{(YY)} - \boldsymbol{\Sigma}_m^{(YXY)}). \tag{4.12}$$

In the conversion process, the converted differential feature vector is determined in the same manner as described in Section 2.3.2.

4.4.3. Gender-dependent modeling

Multiple parallel data sets used in the conventional training method of the MR-GMM consist of singing voice pairs of both male and female singers. To improve modeling accuracy of the MR-GMM on the voice timbre variations, we propose the gender-dependent modeling, inspired by the previous work showing that the voice timbre variations of normal voices caused by aging significantly depend on the gender [68, 69]. Two gender-dependent MR-GMMs are trained separately using the parallel data sets consisting of only male singers or female singers. And then, the singer-dependent MR-GMM for the specific singer is developed by adapting the corresponding gender-dependent MR-GMM to the singer in the same manner as described in Section 2.6.1. Note that not only the representative vectors but also the other parameters, such as the covariance matrices, are different between these two gender-dependent MR-GMMs.

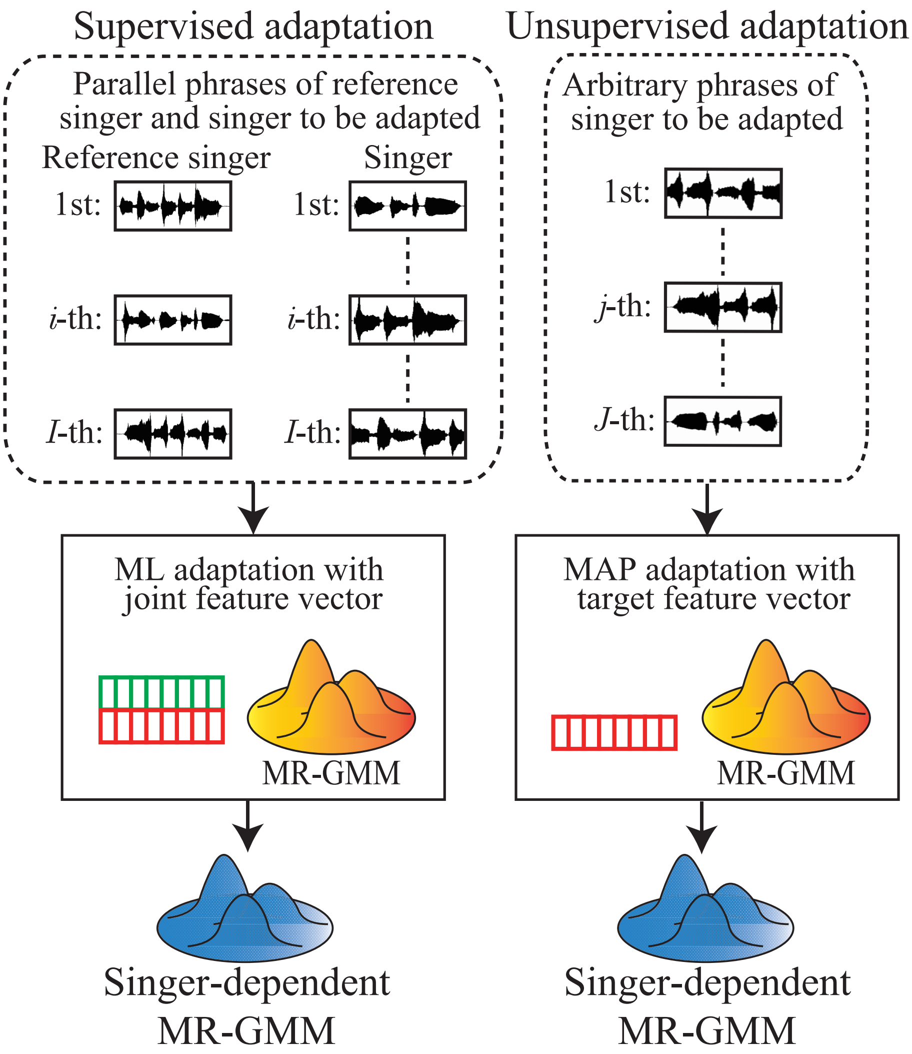

Supervised adaptation
Unsupervised adaptation
Parallel phrases of reference singer and singer to be adapted
Reference singer
Singer
1st:
1st:
i-th:
i-th:
I-th:
I-th:
Arbitrary phrases of singer to be adapted
1st:
j-th:
J-th:
ML adaptation with joint feature vector
MR-GMM
MAP adaptation with target feature vector
MR-GMM
Singer-dependent MR-GMM
Singer-dependent MR-GMM

Figure 4.6.: Adaptation process of perceived age control based on singer-dependent MR-GMM.

4.4.4. Unsupervised adaptation technique for an arbitrary singer

To make it possible to reduce the amount of singing voices and also accept arbitrary phrases used as the adaptation data to develop the singer-dependent MR-GMM, we propose an unsupervised adaptation technique based on the MAP estimation. Figure 4.6 shows the conventional and proposed methods for developing the singer-dependent MR-GMM.

As the prior distribution for the MAP adaptation, the following Gaussian distribution is employed:

$$P\left(\boldsymbol{\mu}|\lambda^{(pri)}\right) = \prod_{m=1}^{M} \mathcal{N}\left(\boldsymbol{\mu}_m; \boldsymbol{\mu}_m^{(pri)}, \boldsymbol{\Sigma}_m^{(pri)}\right), \tag{4.13}$$

where $\lambda^{(pri)}$ is a model parameter set consisting of the mean vectors $\boldsymbol{\mu}^{(pri)} = \left\{\boldsymbol{\mu}_1^{(pri)}, \cdots, \boldsymbol{\mu}_M^{(pri)}\right\}$ and the covariance matrices $\boldsymbol{\Sigma}^{(pri)} = \left\{\boldsymbol{\Sigma}_1^{(pri)}, \cdots, \boldsymbol{\Sigma}_M^{(pri)}\right\}$. This model parameter set is trained in advance using a set of the singer-dependent target mean vectors of all pre-stored target singers as follows:

$$\hat{\lambda}^{(pri)} = \underset{\lambda^{(pri)}}{argmax} \prod_{s=1}^{S} P\left(\boldsymbol{\mu}^{(Y)}(s)|\lambda^{(pri)}\right), \tag{4.14}$$

where $\boldsymbol{\mu}^{(Y)}(s) = \left\{\boldsymbol{\mu}_1^{(Y)}(s), \cdots, \boldsymbol{\mu}_M^{(Y)}(s)\right\}$. For the given adaptation data, $\boldsymbol{Y}(k) = [\boldsymbol{Y}_1^\top(k), \cdots, \boldsymbol{Y}_T^\top(k)]^\top$, which denotes a time sequence of the feature vector of the singer k, the MAP adaptation of the MR-GMM is conducted as follows:

$$\begin{aligned}
\hat{\boldsymbol{\mu}}(k) &= \underset{\boldsymbol{\mu}(k)}{argmax}\, P\left(\boldsymbol{\mu}(k)|\lambda^{(pri)}\right)^{\tau} \int P\left(\boldsymbol{X}, \boldsymbol{Y}(k)|\lambda^{(MR)}, \boldsymbol{\mu}(k)\right) d\boldsymbol{X} \\
&= \underset{\boldsymbol{\mu}(k)}{argmax}\, P\left(\boldsymbol{\mu}(k)|\lambda^{(pri)}\right)^{\tau} \prod_{t=1}^{T} \int P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t(k)|\lambda^{(MR)}, \boldsymbol{\mu}(k)\right) d\boldsymbol{X}_t \\
&= \underset{\boldsymbol{\mu}(k)}{argmax}\, P\left(\boldsymbol{\mu}(k)|\lambda^{(pri)}\right)^{\tau} \prod_{t=1}^{T} P\left(\boldsymbol{Y}_t(k)|\lambda^{(MR)}, \boldsymbol{\mu}(k)\right).
\end{aligned} \tag{4.15}$$

where τ is a hyper-parameter controlling the balance between the prior distribution of mean vectors and the marginalized distribution $P\left(\boldsymbol{Y}(k)|\lambda^{(MR)}, \boldsymbol{\mu}(k)\right)$. The MAP estimate is determined using the EM algorithm by maximizing the following auxiliary function:

$$\begin{aligned}
Q(\boldsymbol{\mu}(k), \hat{\boldsymbol{\mu}}(k)) = \tau \sum_{m=1}^{M} \log P\left(\hat{\boldsymbol{\mu}}_m(k)|\lambda^{(pri)}\right) \\
+ \sum_{t=1}^{T} \sum_{m=1}^{M} P\left(m|\boldsymbol{Y}_t(k), \lambda^{(MR)}, \boldsymbol{\mu}_m(k)\right) \\
\log P\left(\boldsymbol{Y}_t(k), m|\lambda^{(MR)}, \hat{\boldsymbol{\mu}}_m(k)\right).
\end{aligned} \tag{4.16}$$

The MAP estimate is given by

$$\begin{aligned}\hat{\boldsymbol{\mu}}_m(k) &= \left\{\tau \boldsymbol{\Sigma}_m^{(pri)^{-1}} + \Gamma_m \boldsymbol{\Sigma}_m^{(YY)^{-1}}\right\}^{-1} \\ &\quad \cdot \left\{\tau \boldsymbol{\Sigma}_m^{(pri)^{-1}} \boldsymbol{\mu}_m^{(pri)} + \boldsymbol{\Sigma}_m^{(YY)^{-1}} \overline{\boldsymbol{Y}}_m(k)\right\}. \end{aligned} \tag{4.17}$$

4.5. Experimental evaluation for voice timbre control

4.5.1. Evaluation of Modified MR-GMM

Experimental condition

In the first experiment, we evaluated the variation of perceived age achieved by the modified MR-GMM. Eight male subjects in their 20s were divided into two groups, and the 16 evaluation singers were divided into two groups so that one group always included one male singer and one female singer in each age group. We changed the perceived age score in Eq. (4.8) into -60, -40, -20, 0, 20, 40 and 60. Subjects were asked to guess the age of each converted singing voice by listening to it in random order.

In the second experiment, we conducted an XAB test on the singer individuality of both the conventional and modified MR-GMMs. Subjects and evaluation singers were separated into two groups in the same manner as the first experiment. We changed the perceived age score in Eq. (4.8) into -60, -30, 30 and 60 in the modified MR-GMM. In the conventional MR-GMM, the perceived age score in Eq. (4.4) was varied $\pm 30, 60$ from the perceived age of each evaluation singer, which was determined by listening to samples of the intra-singer SVC (source/target) in the previous experiment. A pair of songs generated by the modified and conventional MR-GMM of the same singer and variation of the perceived age scores was presented to subjects after presenting the intra-singer SVC (source) as a reference. Then, they were asked which voice sounded more similar to the reference in terms of the singer individuality.

In the final experiment, we evaluated the naturalness of the converted singing voice using a mean opinion score (MOS). Subjects and evaluation singers were the same as in the first experiment. The perceived age score was the same as for the second evaluation. Subjects rated the naturalness of the converted singing voices using a 5-point scale: “5” for excellent, “4” for good, “3” for fair, “2” for poor, and “1” for

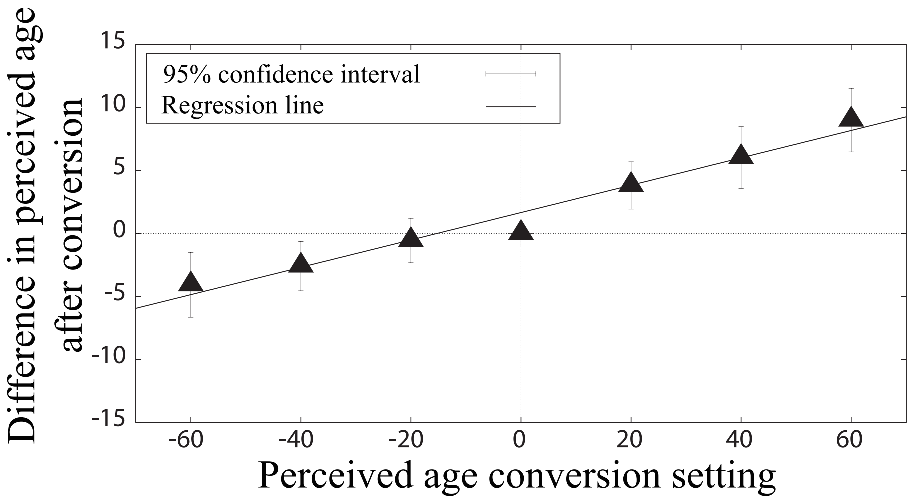


Figure 4.7.: Setting and actual differential in perceived age after conversion.

bad.

In the training of the MR-GMM, we prepared parallel data sets of a single female reference singer in her 20s and 27 male and 27 female singers in their 20s, 30s, 40s and 50s as pre-stored target singers not included in the 16 evaluation singers. The number of training singing voices was 25 in each singer. We used parallel data sets of the reference singer and 16 evaluation singers to update the input mean vectors by Eq. (4.7) for each evaluation singer. The perceived age score for each singer was determined as an average score over 25 singing voices of the singer rated by one male subject in his 20s. The number of mixture components of the MR-GMM was 128 for the spectral envelope and 32 for ACs. The other experimental conditions were the same as Section 4.3.1.

Experimental results

Figure 4.7 indicates the varieties of perceived age in the modified MR-GMM. To change the perceived age score from -60 to 60, the perceived age of the singer was almost linearly varied. In particular, we can see the same tendency as observed in the investigation of segmental features shown in Fig 4.4. The result in Fig 4.4 indicates that the change of observed perceived age from 20 to 60 years old in the horizontal line is about 5 years. This means that modified MR-GMM can appropriately control

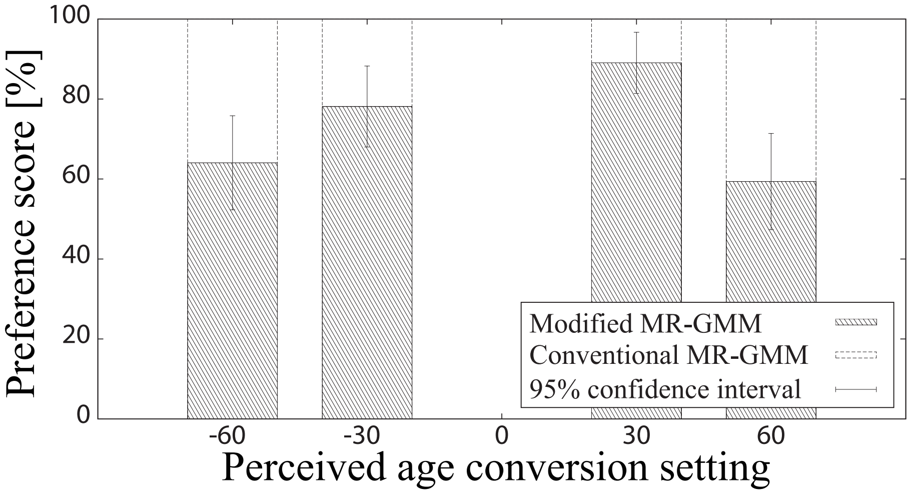


Figure 4.8.: Comparing singer individuality of conventional MR-GMM and modified MR-GMM converted singing voice.

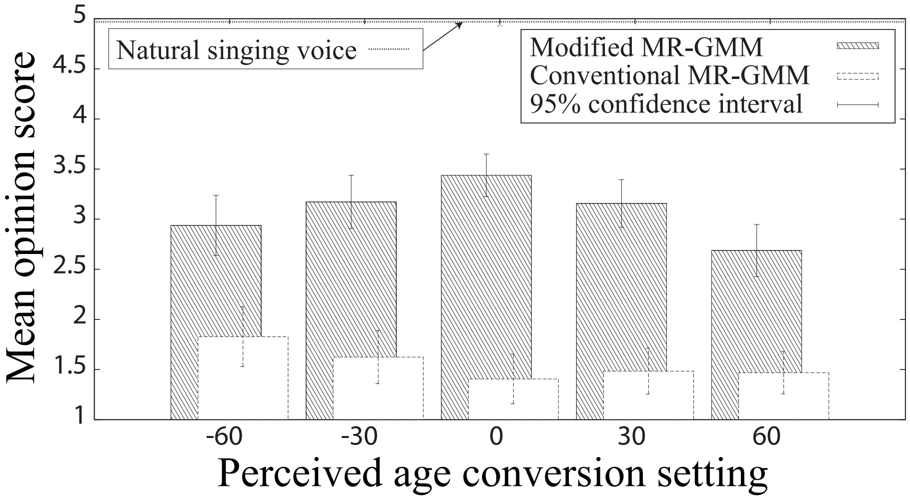


Figure 4.9.: Mean opinion score of conventional MR-GMM and modified MR-GMM.

the perceived age of singing voices.

Figure 4.13 indicates the result of the XAB test for the singer individuality. We can see that as we make larger changes in the perceived age, the preference score of the modified MR-GMM tends to decrease. However, the modified MR-GMM has a higher preference score than the conventional MR-GMM for each setting.

Figure 4.12 indicates the results of MOS test for the naturalness. This figure has

Table 4.4.: Experimental conditions of evaluation of gender-dependent Modified DIFFMR-GMM.

Singing voice database	AIST humming database
Sampling frequency	16 [kHz]
Duration of one phrase	about 20 [s]
The number of training singers	28 males, 28 females
The number of evaluation singers	8 males, 8 females
The number of training data	23 phrases
The number of subjects	8

the same tendency as displayed in Figure 4.13. The modified MR-GMM has a higher MOS than the conventional MR-GMM for each setting. The bias vectors of the modified MR-GMM ($\hat{\boldsymbol{\mu}}_m^{(Y)}$ in Eq. (4.8)) model singing voice characteristics of a single singer (i.e., the source singer). On the other hand, those of the conventional MR-GMM ($\overline{\boldsymbol{\mu}}_m^{(Y)}$ in Eq. (4.8)) model voice characteristics of multiple pre-stored target singers. Therefore, over-smoothing effects of the conventional MR-GMM tend to be larger than those of the modified MR-GMM. Consequently, the naturalness of the singing voices is also improved by using the modified MR-GMM.

These results suggest that 1) the modified MR-GMM enables to control the perceived age of singing voices relatively well, 2) the modified MR-GMM enables to retain the singer individuality better than the conventional MR-GMM during the perceived age control, and 3) the modified MR-GMM also generates better quality of converted singing voices compared with the conventional MR-GMM.

4.5.2. Evaluation of gender-dependent Modified DIFFMR-GMM

Experimental condition

Table 4.4 indicates a simple description of the experimental conditions. We used the AIST humming database [67] consisting of phrases of songs with Japanese lyrics sung by Japanese male and female amateur singers in their 20s, 30s, 40s, and 50s. The sampling frequency was set to 16 kHz. The 1st through 24th mel-cepstral coefficients extracted by STRAIGHT analysis [12] were used as spectral features. As the source excitation features, we used F_0 and aperiodic components in five frequency bands, i.e.,

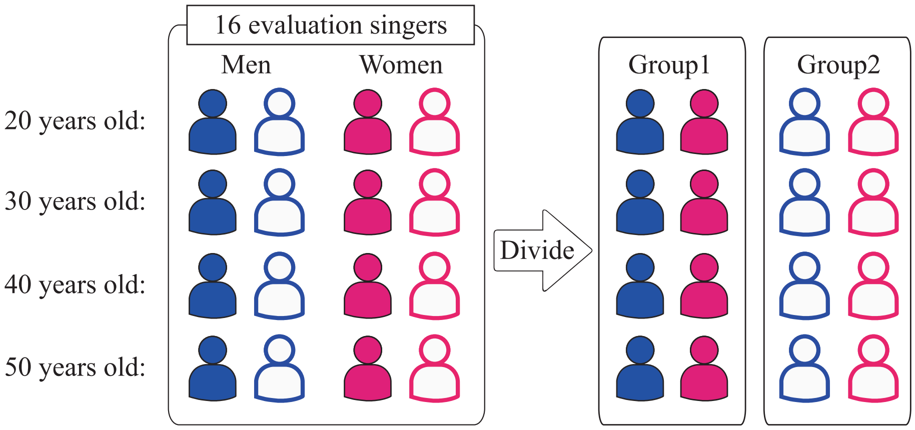


Figure 4.10.: Method for dividing 16 evaluation singers into two groups.

0–1, 1–2, 2–4, 4–6, and 6–8 kHz, which were also extracted by STRAIGHT analysis [49]. The frame shift was 5 ms. The mel log spectrum approximation filter [40] was used as the synthesis filter in both the conventional waveform generation with vocoder and the proposed direct waveform modification.

In the training of the gender-independent MR-GMM, we used parallel data sets of a female reference singer in her 20s and 56 pre-stored target singers including 28 males and 28 females in their 20s, 30s, 40s and 50s. In the training of the gender-dependent MR-GMMs, we separately used a female and male reference singer in their 20s and 28 male or 28 female pre-stored target singers. Each singer sang 23 phrases, where the duration of each phrase was approximately 20 seconds. The number of mixture components of each MR-GMM was 128 for the spectral feature and 64 for the aperiodic components. We have developed the singer-dependent MR-GMMs for 16 singers consisting of two male and two female singers in each age group (the 20s, 30s, 40s, and 50s), who were not included in the pre-stored target singers, and conducted voice timbre control evaluations for these singers. We used P039 as an evaluation phrase. The perceived age score for each singer was determined as an average score of the singer rated by 8 subjects in their 20s [70].

To examine the effectiveness of two proposed techniques, the gender-dependent modeling, and the direct waveform modification, singing voices converted by the following three methods were evaluated:

- SVC (GI): converted with the gender-independent Modified MR-GMM,

- SVC (GD): converted with the gender-dependent Modified MR-GMM,
- DIFFSVC (GD): converted with the gender-dependent Modified DIFFMR-GMM.

The converted singing voice samples were generated by settings of the perceived age score differential to -60, -30, 0, 30, and 60. The number of training phrases for the development of the singer-dependent MR-GMM was 23 in each singer. Figure 4.10 indicates a method of dividing 16 evaluation singers into two groups. The 16 evaluation singers were divided into two groups so that one group always included one male singer and one female singer in each age group. Each subject was assigned one evaluation singer group in each evaluation in order to evaluate the evaluation singers of both genders and all age groups.

First, we evaluated perceived age controllability. The number of the converted singing voice of an evaluation singer was 15. Each subject evaluated the converted singing voices of 120 phrases from only one group of the evaluation singers. Subjects were asked to assign the perceived age to each converted singing voice sample by listening to it in random order.

In the second experiment, we evaluated the quality of the converted singing voice using a mean opinion score (MOS). Each subject evaluated the natural and converted singing voices of evaluation singers. The number of evaluation phrases in each subject is 128. The subjects rated the quality of the converted singing voice using a 5–point scale: “5” for excellent, “4” for good, “3” for fair, “2” for poor, and “1” for bad.

In the final experiment, we conducted an XAB test on the singer individuality to compare the conventional method SVC (GI) and the proposed method DIFFSVC (GD). The evaluation singers were separated into two groups and each subject evaluated the converted singing voices from only one group in the same manner as the first experiment. A pair of singing voices converted by SVC (GI) and by DIFFSVC (GD) for the same singer with the same setting of the perceived age score differential was presented to the subjects after presenting the natural singing voice as a reference. Then, they were asked which singing voice sounded more similar to the reference in terms of the singer individuality. The number of evaluation pairs in each subject is 40.

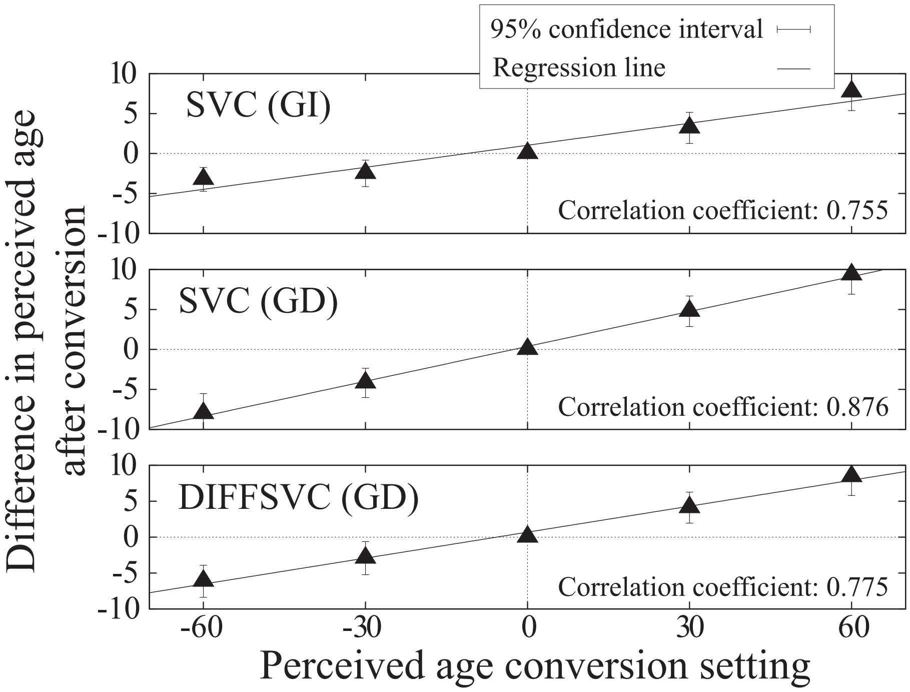


Figure 4.11.: Experimental result on perceived age controllability.

Experimental result

Figure 4.11 shows the relationship between the perceived age differentials given to the system to generate the converted singing voices and their perceived ages actually evaluated by the listeners. We can see that using the proposed gender-dependent models (SVC (GD) and DIFFSVC (GD)), the perceived age varies more linearly according to a change of the settings of the perceived age differential from -60 to 60 compared with the conventional gender-independent model (SVC (GI)). Moreover, a range of the perceived age of the converted singing voice becomes wider by using SVC (GD) and DIFFSVC (GD) compared with SVC (GI). These results indicate that voice timbre variations caused by the perceived age depend on the gender in singing voices and they are well modeled by using the proposed gender-dependent modeling technique.

Figure 4.12 indicates the results of the opinion test on the quality. We can see that DIFFSVC (GD) tends to significantly improve the quality of the converted singing voices compared with SVC (GI) and SVC (GD). Although the quality is greatly degraded in the conventional method SVC (GI) as the perceived age score differential

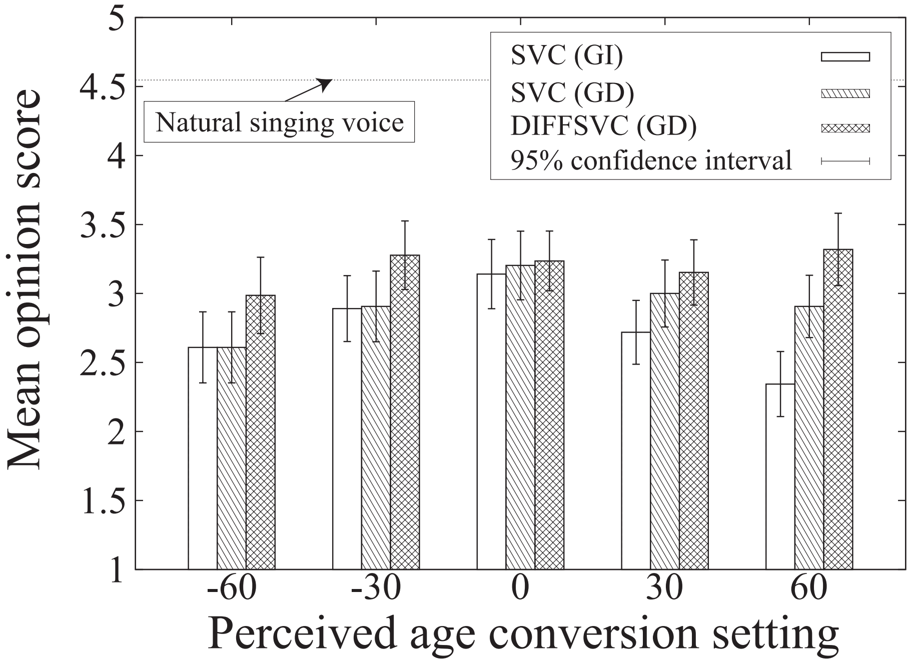


Figure 4.12.: Mean opinion score of sound quality.

is set to larger or smaller values, this quality degradation is effectively alleviated by the proposed method DIFFSVC (GD) because the DIFSVC (GV) method can avoid the errors caused by spectrum parameterization and excitation generation. In comparison between SVC (GD) and SVC (GI), the speech quality of SVC (GD) is improved compared with that of SVC (GI) as the perceived age score differential is set to higher values (+30, +60). On the other hand, in terms of setting lower values (−60, −30), we can see that there is no significant difference between these methods. As shown in Figure 4.11, the perceived age differential achieved by SVC (GI) tends to be smaller than that by SVC (GD) when setting the perceived age score differential to −60. This result implies that the resulting acoustic changes by SVC (GI) are smaller than those by SVC (GD) under such a setting, also making the quality degradation in SVC (GI) smaller. Even in such an unfair condition, SVC (GD) causes no quality degradation compared with SVC (GI).

Figure 4.13 indicates the result of the XAB test on the singer individuality. DIFFSVC (GD) better or equally retains singer individuality in any perceived age setting compared with the conventional method SVC (GI). We can see that as a change of the

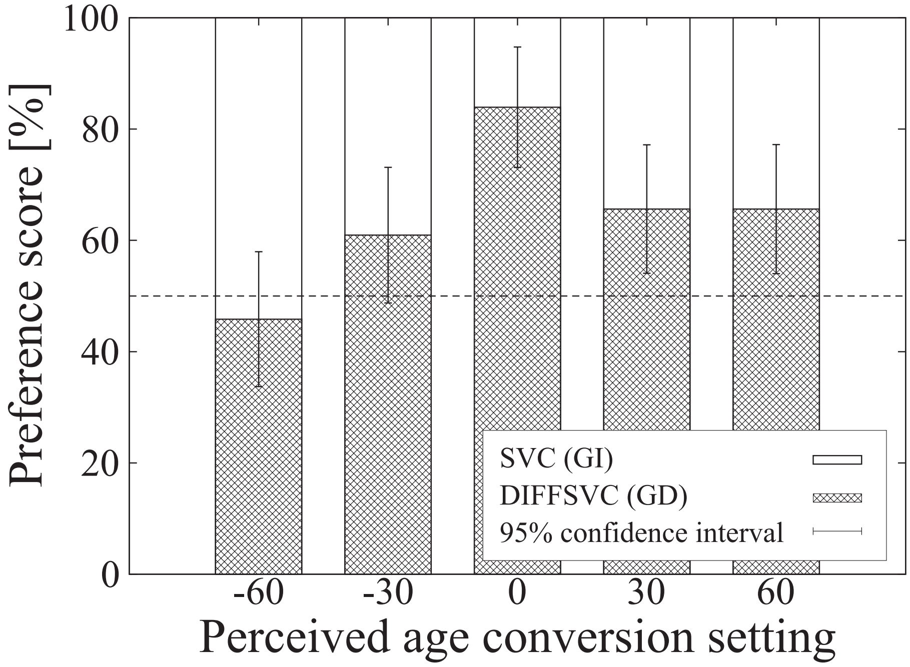


Figure 4.13.: Preference score on singer individuality.

perceived age differential setting is larger, the difference between DIFFSVC (GD) and SVC (GI) becomes smaller. In particular, no difference is observed between them when setting the perceived age differential to −60 while the significant difference is still observed when setting it to 60. It is expected that this result is also caused by the resulting acoustic changes by SVC (GI) is smaller than SVC (GI) when setting the perceived age differential to −60 as mentioned above.

These results suggest that 1) the gender-dependent modeling technique is effective for improving the perceived age controllability, and 2) the direct waveform modification technique with spectral differential significantly improves the quality of the converted singing voice.

Although the proposed method DIFFSVC (GD) makes it possible to control the perceived age with higher speech quality compared with the conventional method SVC (GI) and SVC (GD) in Figures 4.12 and 4.13, there still remains the speech quality degradation compared with the natural singing voice. It is expected that this degradation is caused by insufficient modeling accuracy of the perceived age variations using the gender-dependent MR-GMM. Therefore, it is worthwhile to further improve the modeling accuracy.

4.5.3. Evaluation of unsupervised adaptation

Experimental condition

In this evaluation, we varied the number of the adaptation phrases as 1, 6, 12, and 22 in order to evaluate the effectiveness of the proposed unsupervised adaptation technique. The adaptation phrases are selected in order from the beginning of the index of singing voice database. In this evaluation, the ML estimation with parallel phrases was used as the supervised adaptation and the MAP estimation with only phrases of each evaluation singer was used as the unsupervised adaptation. The hyper-parameter τ for the MAP adaptation was manually set to 3.0 in the subjective evaluations.

First, we evaluated the modeling accuracy of the singer-dependent MR-GMMs developed with the adaptation approaches using Mahalanobis distance of their mean vectors to those of the singer-dependent MR-GMMs developed with the conventional supervised approach using 22 parallel phrases in each singer, which is calculated as

$$D(i) = \frac{1}{L} \sum_{l=1}^{L} \sum_{m=1}^{M} \alpha_m \left(\boldsymbol{\mu}_m^{(22)}(l) - \hat{\boldsymbol{\mu}}_m^{(i)}(l) \right)^\top \Sigma_m^{(YY)^{-1}} \left(\boldsymbol{\mu}_m^{(22)}(l) - \hat{\boldsymbol{\mu}}_m^{(i)}(l) \right), \tag{4.18}$$

where L denotes the number of evaluation singers. $\hat{\boldsymbol{\mu}}_m^{(i)}(l)$ denotes the adapted singer-dependent MR-GMM for the evaluation singer l using his/her i phrases in the unsupervised adaptation or i parallel phrases in the supervised adaptation. Note that the mean vectors of the singer-dependent MR-GMM used as a target $\boldsymbol{\mu}_m^{(22)}(l)$ in this distance calculation is equivalent to those determined using the supervised ML adaptation using 22 parallel phrases.

In the second experiment, we evaluated the conversion accuracy using the mel-cepstrum distortion as an evaluation metric in the different settings of the hyper-parameter τ. The mel-cepstrum distortion was calculated as follows:

$$\text{Mel-CD } [dB] = \frac{10}{\ln 10} \sqrt{2 \sum_{d=1}^{24} \left(mc_d^{(22)} - \hat{mc}_d^{(i)} \right)^2}, \tag{4.19}$$

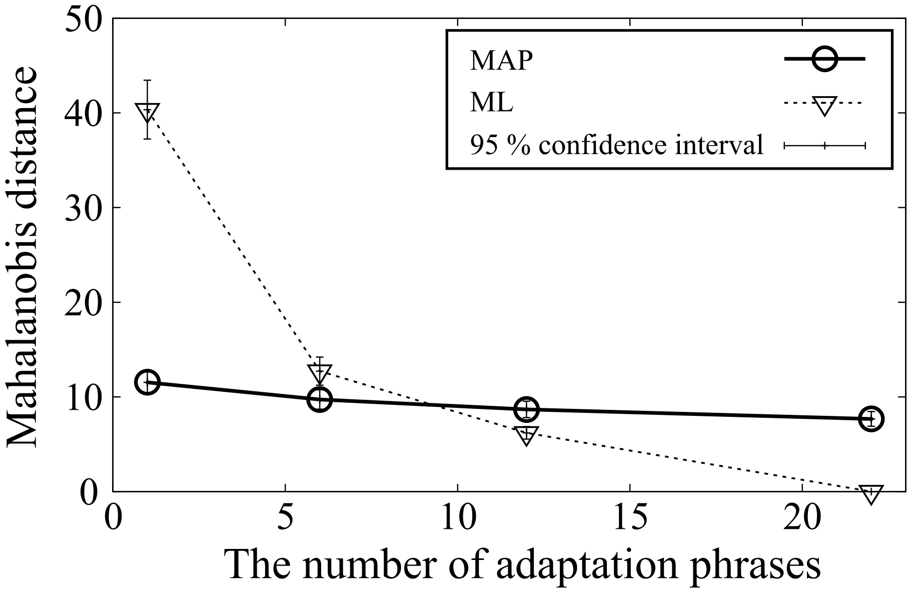


Figure 4.14.: Mahalanobis distance as a function of the number of adaptation phrases.

where $mc_d^{(22)}$ denotes the mel-cepstrum coefficients analyzed from the converted singing voice generated with the singer-dependent MR-GMM developed with the supervised ML adaptation using 22 parallel phrases, and $\hat{mc}_d^{(i)}$ denotes those developed with the unsupervised MAP adaptation using i phrases. The setting of the hyper-parameter τ is varied from 0, 1, 3, 6, 12, to 24. Note that the setting of $\tau = 0$ corresponds to the unsupervised ML adaptation.

In the third experiment, we evaluated the perceived age controllability. The number of adaptation phrases was set to 1 and 6. The 16 evaluation singers were divided into four groups. Each subject evaluated the converted singing voices from only one group of the evaluation singers. Subjects were asked to assign the perceived age to each converted singing voice in one group of the evaluation singers by listening to it in random order. The number of evaluation samples in each subject was 48.

In the final experiment, we evaluated the quality of the converted singing voice using an opinion test. The number of subjects and evaluation singers was the same as in the second experiment. The subjects evaluated the quality of the converted singing voices in the same manner as described in 4.5.2.

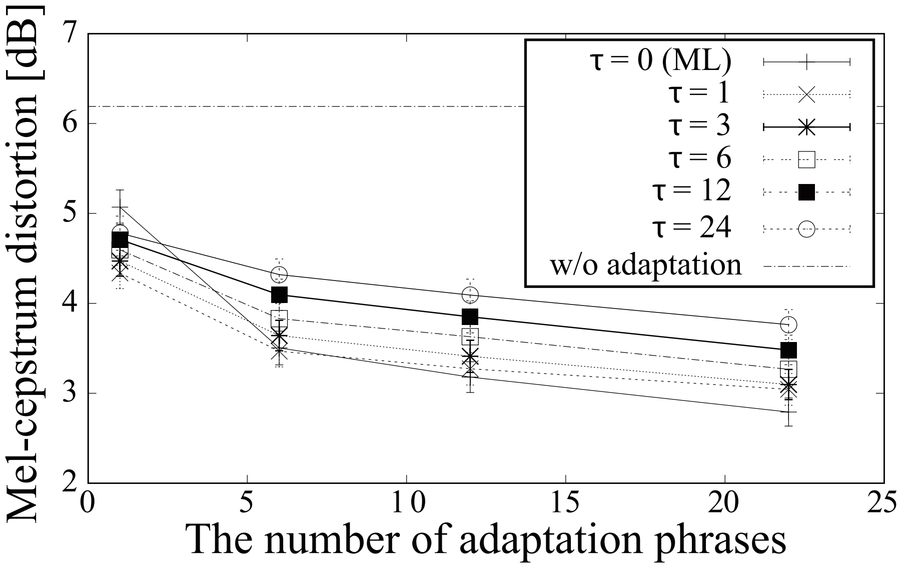


Figure 4.15.: Mel-cepstrum distortion as a function of the number of adaptation phrases and hyper-parameters settings.

Experimental result

Figure 4.14 indicates the Mahalanobis distances as a function of the number of adaptation phrases. The distance when using 1 parallel phrase in the ML adaptation is very large. On the other hand, the distance using 1 phrase in the MAP adaptation is significantly lower than it. In the ML adaptation, it is necessary to use 6 or more parallel phrases to reduce the distance as small as in the MAP adaptation.

Figure 4.15 shows the mel-cepstrum distortion as a function of the number of adaptation phrases in each hyper-parameter setting. We can see that the unsupervised adaptation using either ML or MAP is effective. The unsupervised ML adaptation ($\tau = 0$) causes significantly large degradation when using only one adaptation phrase. On the other hand, such a degradation is effectively alleviated by using the proposed MAP adaptation. We can also see that performance of the proposed MAP adaptation is affected by the hyper-parameter setting, and relatively good performance is achieved by setting the hyperparameter to a small value.

Figure 4.16 shows the experimental result on the perceived age controllability. We can see that the MAP adaptation using only 1 phrase has higher controllability compared with the ML adaptation in 1 parallel phrase and its controllability is similar to

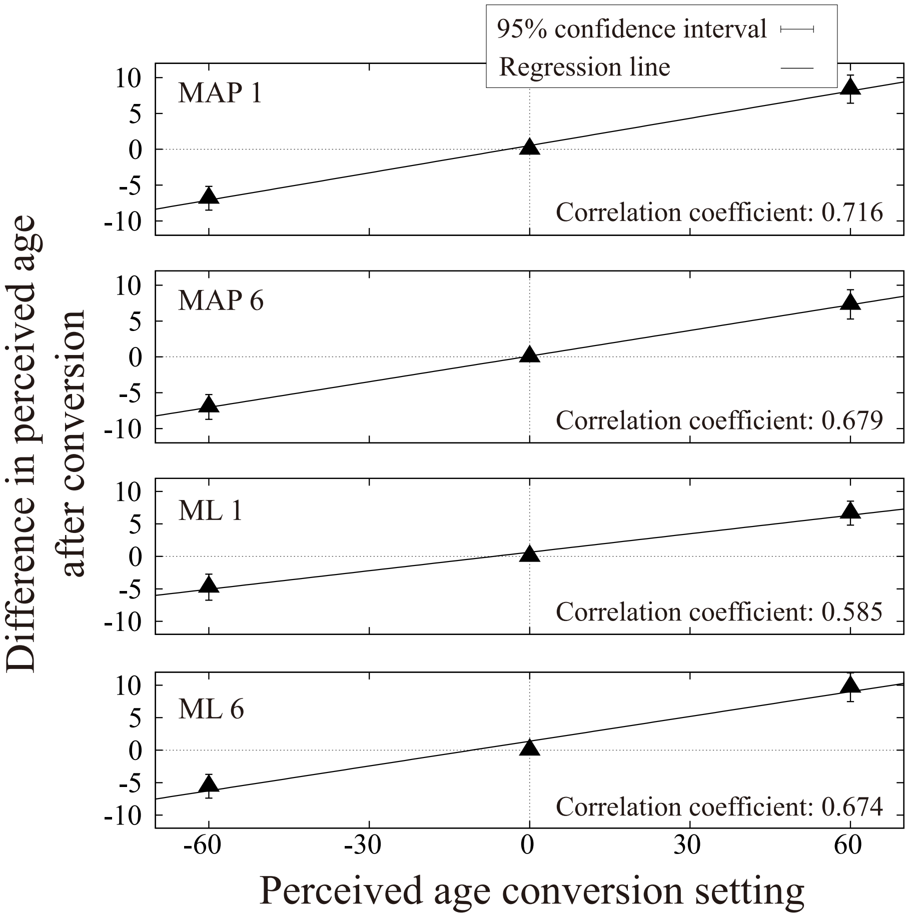


Figure 4.16.: Experimental result on perceived age controllability of the adapted MR-GMMs.

that of the MAP adaptation using 6 phrases and that of the ML adaptation using 6 parallel phrases. This tendency is consistent with that observed in the previous objective evaluation shown in Figure 4.14. Moreover, comparing to the result described in Figure 4.11, we can see that the proposed MAP adaptation method using only 1 phrase achieves similar controllability to the conventional method using 22 parallel phrases.

Figure 4.17 indicates the results of the opinion test on the speech quality. We can see that there is no significantly large quality difference between the MAP adaptation and the ML adaptation. We can also see that the quality of the converted singing voice tends to degrade if using only 1 phrase. This quality degradation is alleviated by increasing the number of adaptation phrases to 6 and the resulting quality reaches to

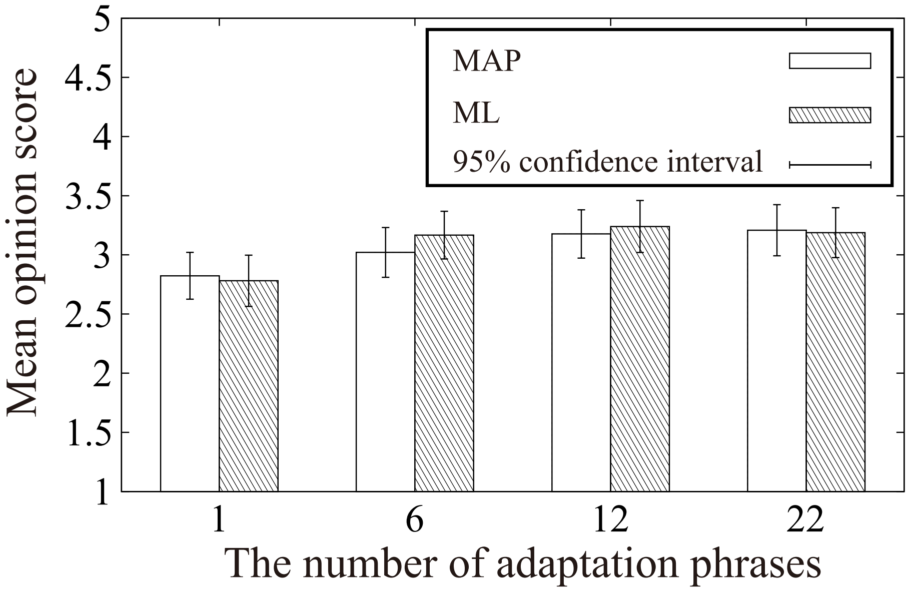


Figure 4.17.: Mean opinion score of speech quality depending on the number of adaptation phrases.

that of the conventional method using 22 parallel phrases.

These results suggest that 1) the MAP adaptation outperforms the ML adaptation when a few phrases are available, and 2) the MAP adaptation by using only a small number of arbitrary phrases (e.g., 6 phrases) achieves almost the same controllability and quality of the converted singing voice as in the conventional method that needs a larger number of parallel phrases (e.g., 22 phrases).

4.6. Summary

In this chapter, in order to control voice timbre of singing voice based on perceptually understandable voice timbre expression words while retaining singer identity, we have investigated acoustic features affecting on the perceived age and proposed voice timbre control technique with the Modified MR-GMM and the gender-dependent Modified DIFFMR-GMM.

Sections 4.2 and 4.3: These sections illustrated the investigation of the acoustic features affecting on the perceived age in order to asses effects of each acoustic feature. The experimental results showed that 1) the perceived age of singing voices corresponds relatively well to the actual age of the singer, 2) prosodic features have a larger effect on the perceived age than spectral features, 3) the individuality of a singer is influenced more heavily by segmental features than prosodic features.

Sections 4.4 and 4.5: In these sections, in order to control voice timbre of the singer based on the perceived age while retaining singer identity, we have proposed a statistical voice timbre control technique based on Modified MR-GMM. The experimental results confirmed that the proposed technique based on the Modified MR-GMM makes it possible to achieve to control perceived age of the source singing voice while not having an adverse effect on singer identity. Moreover, to improve the controllability of the perceived age and source quality of the converted singing voice, we have proposed a perceived age control technique based on direct waveform modification using time-variant spectral differential with gender-dependent Modified DIFFMR-GMM. The experimental results indicated the perceived age control technique based on the gender-dependent Modified DIFFMR-GMM achieved higher controllability and sound quality compared with the perceived age control technique based on Modified MR-GMM. Furthermore, to develop the singer-dependent model for an arbitrary source singer easily, we have proposed a rapid adaptation technique based on MAP adaptation. The experimental results have demonstrated that the unsupervised adaptation technique makes it possible to develop the singer-dependent model for the arbitrary source singer even when parallel phrases are not available.

5. Real-time VC systems via statistical waveform modification

5.1. Introduction

In speech communication, speakers can transmit their speech consisting of linguistic and non-linguistic information utilizing their speech production mechanism. Voice timbre is one of the most important non-linguistic information for listeners to identify the speakers' individuality. In order to convert an individuality of a source speaker into that of a target speaker, several voice conversion (VC) techniques such as Gaussian process regression [19, 20] and deep neural network [21–23] have been proposed. Although these techniques achieved a conversion of voice timbre with higher conversion accuracy compared with VC based on Gaussian mixture model (GMM) [8, 10], it is difficult to utilize these techniques into a real-time conversion system because these techniques are difficult to convert voice timbre in real-time.

A real-time VC system has some possibility of growing a valuable interactive communication tool. For example, it is expected that the real-time VC system makes it possible to expand our speech expressions in various conditions such as speech disorder, pronunciation correction of a non-native language, impersonation and so on. If the speakers freely produced various voice timbre using the real-time VC system, it would open up a new speech communication style.

In order to make it possible to implement such that VC system, in this chapter, we propose real-time statistical waveform modification systems using a low-delay conversion technique [25]. Although several real-time conversion systems such as lip-synching into speech waveform [71], a timbre of a musical instrument [72], pitch of singing voice [73] have been proposed as the real-time conversion systems when user's input is given, these techniques are not possible to convert a speaker individuality of a user into that of another user. In Chapters 3 and 4, we have proposed VC and voice timbre control techniques via statistical waveform modification to improve the sound quality of the converted voice. However, it is not possible to directly apply these techniques into the low-delay conversion framework. In this chapter, in order to apply the statistical waveform modification techniques into real-time conversion systems, we propose following techniques: 1) parameter transformation technique for the low-delay conversion and 2) frame-based global variance (GV) post-filter to alleviate the over-smoothing effect of the converted feature trajectory.

This chapter is organized as follows. System overview consisting of our implemented VC and voice timbre control systems is illustrated in Section 5.2. System components consisting of a parameter transformation technique for the low-delay conversion and frame-based GV post-filter for statistical waveform modification are described in Section 5.3. The experimental evaluations are described in Section 5.4. This chapter is summarized in Section 5.5.

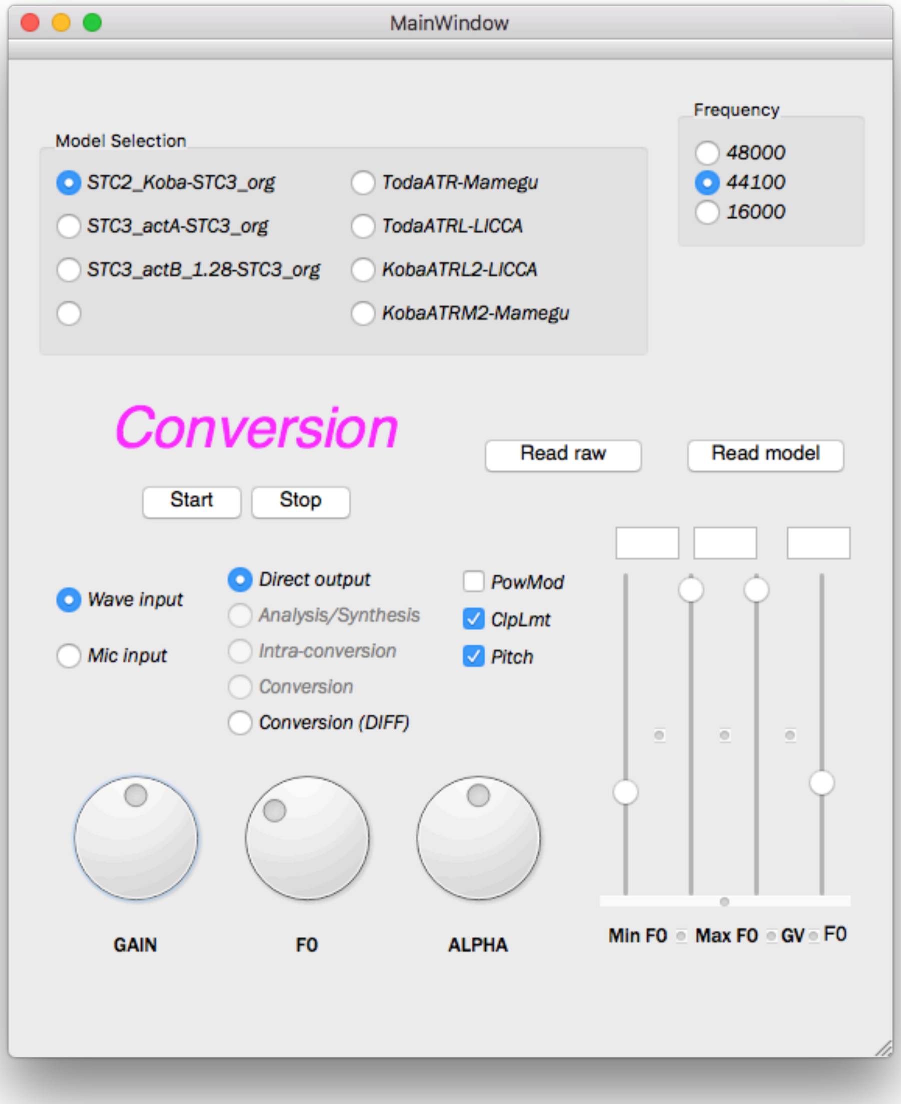
5.2. System overview

5.2.1. Real-time VC system based on statistical waveform modification

Figure 5.1 illustrates a graphical user interface of the real-time VC system. In this interface, it is possible to immediately change a conversion model between the source and target speakers by clicking “Model Selection” push buttons. In order to transform F_0 of an input voice, we install F_0 control sliders.

5.2.2. Real-time voice timbre control system based on statistical waveform modification

Figure 5.2 illustrates a graphical user interface of the real-time perceived age control system. In this interface, it is possible to control the perceived age of an input singing voice by manipulating a slider in accordance with the perceived age to “up” or “down”. The conversion begins when pushing a “start” button after choosing pre-stored voice samples. In this system, the user can change the perceived age slider while confirming the output converted voice. Therefore, it is easy for the user to find the ideal voice



Model Selection

● STC2_Koba-STC3_org
○ STC3_actA-STC3_org
○ STC3_actB_1.28-STC3_org
○
○ TodaATR-Mamegu
○ TodaATRL-LICCA
○ KobaATRL2-LICCA
○ KobaATRM2-Mamegu

Frequency

○ 48000
● 44100
○ 16000

● Wave input
○ Mic input

● Direct output
○ Analysis/Synthesis
○ Intra-conversion
○ Conversion
○ Conversion (DIFF)

□ PowMod
☑ ClpLmt
☑ Pitch

Figure 5.1.: Graphical user interface of a real-time VC system.

timbre. Moreover, to confirm the reference perceived age, the system prepares several representative voice timbre in each age.

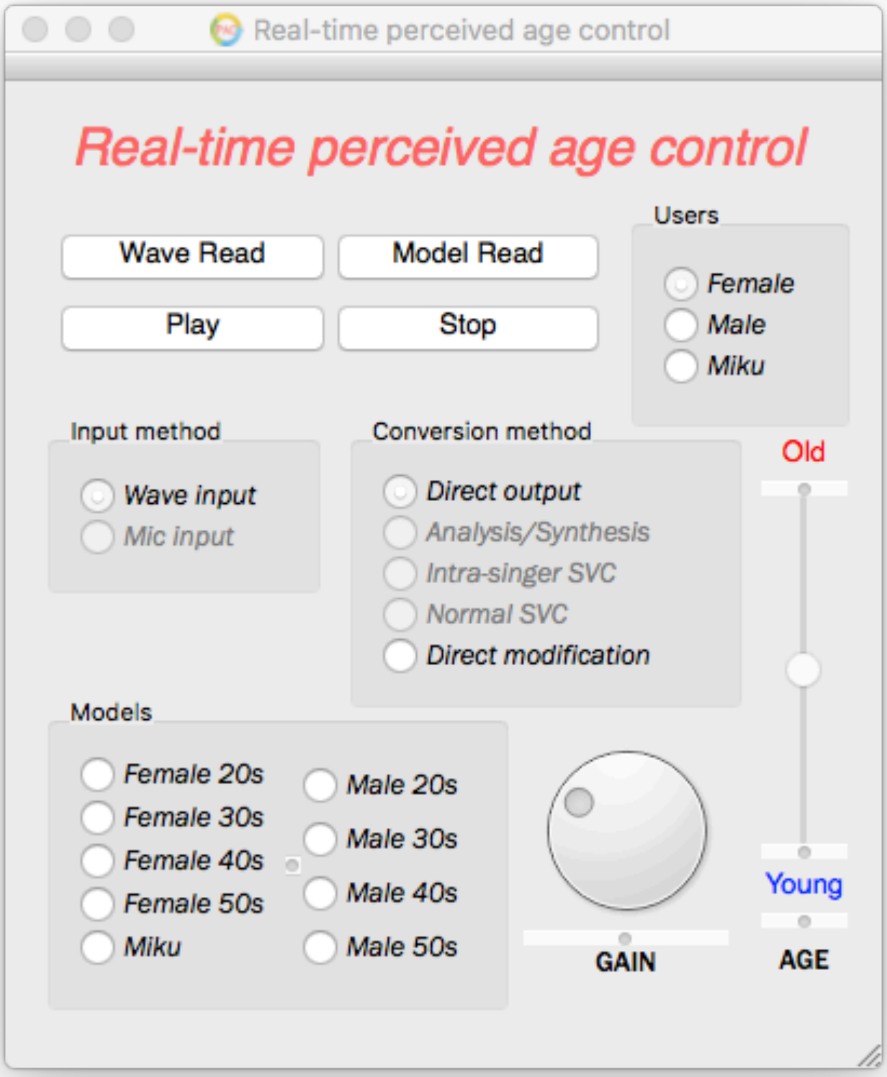


Figure 5.2.: Graphical user interface of a real-time voice timbre control system based on perceived age.

5.3. System components

In this section, we propose following two techniques to implement real-time VC and voice timbre control systems: 1) parameter transformation technique for low-delay statistical waveform modification and 2) frame-based GV post-filtering for statistical waveform modification to alleviate the over-smoothing effect of the converted feature trajectory.

5.3.1. Model parameter transformation for low-delay statistical waveform modification

In the conventional low-delay conversion algorithm [25], to reduce computational costs of the feature extraction, source spectral envelope of an input waveform is extracted based on a simple spectral feature extraction method based on fast Fourie transform (FFT) using fixed window and liftering. And also, spectral envelope extracted based on STRAIGHT analysis [12] is used as a target spectral envelope in order to achieve the higher sound quality of the converted voice compared with the spectral envelope extracted based on FFT using fixed window and liftering. Therefore, it is not possible to directly apply the joint probability density function of a differential GMM (DIFFGMM) modeled using joint source and target spectral features extracted using STRAIGHT analysis as the same manner as described in Chapter 3 into the low-delay conversion algorithm. In this section, in order to estimate the joint probability density function for the real-time statistical waveform modification systems, we propose parameter transformation technique.

Let $\dot{\boldsymbol{X}}_t = [\dot{\boldsymbol{x}}_t^\top, \Delta\dot{\boldsymbol{x}}\top_t]^\top$, $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$, and $\boldsymbol{Y} = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$ denote $2D$-dimensional joint static and dynamic feature vectors of the source extracted based on the simple spectral envelope extraction and the source and target extracted based on STRAIGHT analysis, respectively, where $\top$ denotes the transposition of the vector. These feature vectors consist of D-dimensional static feature vectors $\dot{\boldsymbol{x}}_t$, $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ and their dynamic feature vectors $\Delta\dot{\boldsymbol{x}}_t$, $\Delta\boldsymbol{x}_t$ and $\Delta\boldsymbol{y}_t$ at frame t, respectively. Their joint probability density modeled by the GMM is given by

$$P\left(\dot{\boldsymbol{X}}_t, \boldsymbol{X}_t, \boldsymbol{Y}_t | \boldsymbol{\lambda}^{(\dot{X}XD)}\right) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix} \dot{\boldsymbol{X}}_t \\ \boldsymbol{X}_t \\ \boldsymbol{Y}_t \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu}_m^{(\dot{X})} \\ \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_m^{(\dot{X}\dot{X})} & \boldsymbol{\Sigma}_m^{(\dot{X}X)} & \boldsymbol{\Sigma}_m^{(\dot{X}Y)} \\ \boldsymbol{\Sigma}_m^{(X\dot{X})} & \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(Y\dot{X})} & \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix}\right), \tag{5.1}$$

where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. The total number of mixture components is M. The mixture component index is m. $\boldsymbol{\lambda}$ is a GMM parameter set consisting of the mixture-component weight α_m, the mean vector $\boldsymbol{\mu}_m$, and the covariance matrix $\boldsymbol{\Sigma}_m$ of the m-th mixture component. Note that we use single-path training to model this joint probability density.

In order to estimate joint probability density function for the low-delay conversion systems, we define following transformation matrix:

$$\boldsymbol{A}^{(RT)} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{-I} & \boldsymbol{I} \end{bmatrix}. \tag{5.2}$$

Let $\boldsymbol{D}_t = \left[\boldsymbol{d}_t^\top, \Delta\boldsymbol{d}_t^\top\right]^\top$ denote the static and dynamic differential feature vector, where $\boldsymbol{d}_t = \boldsymbol{y}_t - \boldsymbol{x}_t$. Applying to the transformation matrix $\boldsymbol{A}^{(RT)}$ into the joint probability density function in Eq. 5.1, the joint probability density function of the low-delay statistical waveform modification is derived as follows:

$$P\left(\dot{\boldsymbol{X}}_t, \boldsymbol{D}_t | \boldsymbol{\lambda}^{(\dot{X}D)}\right) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix} \dot{\boldsymbol{X}}_t \\ \boldsymbol{D}_t \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu}_m^{(\dot{X})} \\ \boldsymbol{\mu}_m^{(D)} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_m^{(\dot{X}\dot{X})} & \boldsymbol{\Sigma}_m^{(\dot{X}D)} \\ \boldsymbol{\Sigma}_m^{(D\dot{X})} & \boldsymbol{\Sigma}_m^{(DD)} \end{bmatrix}\right), \tag{5.3}$$

$$\boldsymbol{\mu}_m^{(D)} = \boldsymbol{\mu}_m^{(Y)} - \boldsymbol{\mu}_m^{(X)}, \tag{5.4}$$

$$\boldsymbol{\Sigma}_m^{(\dot{X}D)} = \boldsymbol{\Sigma}_m^{(D\dot{X})^\top} = \boldsymbol{\Sigma}_m^{(\dot{X}Y)} - \boldsymbol{\Sigma}_m^{(\dot{X}X)}, \tag{5.5}$$

$$\boldsymbol{\Sigma}_m^{(DD)} = \boldsymbol{\Sigma}_m^{(XX)} + \boldsymbol{\Sigma}_m^{(YY)} - \boldsymbol{\Sigma}_m^{(XY)} - \boldsymbol{\Sigma}_m^{(YX)}. \tag{5.6}$$

5.3.2. Frame-based GV post-filter for low-delay statistical waveform modification

For DIFFVC considering GV described in 3.2, it is necessary to iteratively estimate spectral feature differential while considering the GV of the converted feature trajec-

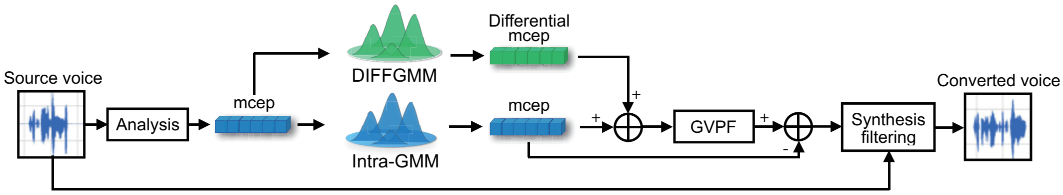


Figure 5.3.: A conversion flow of the low-delay statistical waveform modification considering frame-based GV.

tory. If the source spectral feature is analyzed using STRAIGHT analysis [12] likewise the off-line conversion, it is possible to estimate the converted spectral feature trajectory by working out the sum of the source spectral feature and the differential. On the other hand, in the low-delay conversion framework, simple spectral feature extraction process consisting of FFT using fixed window width and liftering is performed. Therefore, the conversion accuracy of the spectral feature differential is degraded because spectral envelope modeling accuracy of the simple spectral feature extraction is usually worth compared with that of STRAIGHT analysis. Consequently, this degradation arises the sound quality degradation of the converted voice.

In this section, in order to alleviate this sound quality degradation, we propose a frame-based GV post-filtering technique for low-delay statistical waveform modification using intra-speaker VC technique. Figure 5.3 indicates a conversion flow of the low-delay statistical waveform modification considering frame-based GV. Using intra-speaker VC, it is possible to approximately estimate the source spectral feature extracted using STRAIGHT from the source spectral feature extracted simple spectral feature extraction.

Let $\hat{d}_{t,d}$ and $\hat{x}_{t,d}$ denote d dimensional converted static differential feature vector without considering GV and converted static feature vector based on intra-speaker VC. The frame-based GV post-filtering process follows:

$$\hat{d}^{(GV)}_{t,d} = \mu_d^{(v)\frac{1}{2}} \bar{\mu}_d^{(v)-\frac{1}{2}} \left(\hat{x}_{t,d} + \hat{d}_{t,d} - \bar{y}_d\right) + \bar{y}_d - \hat{x}_{t,d}, \tag{5.7}$$

where $\mu_d^{(v)}$ is GV of d-th dimensional static feature vector of the target speaker. $\bar{\mu}_d^{(v)}$ and $\bar{y}_d$ denote GV and mean vector of d-th dimensional converted static feature vector without considering GV, which are previously calculated using DIFFVC without considering GV. Although the proposed post-filtering technique is necessary to convert double at the same time because it requires both the converted static feature vectors of intra-speaker VC and the converted static differential feature vector of DIFFVC, these conversions are capable of converting in real-time. Note that we also suppress a conversion using spectral feature differential by padding to zero at the unvoiced frame.

5.4. Experimental evaluation

5.4.1. Experimental condition

We used singing voices of 21 Japanese traditional songs, which were divided into 152 phrases, where the duration of each phrase was approximately 8 seconds. Amateur singers including 3 males and 3 females sang these phrases. The sampling frequency was set to 16 kHz. The frame shift was 5 ms. STRAIGHT [12] was used to extract spectral envelopes, which were parameterized to the 1-24th mel-cepstral coefficients as spectral features. The mel log spectrum approximation (MLSA) filter [40] was used as the synthesis filter. We used 80 phrases for the GMM training and the remaining 72 phrases were used for evaluation. GMMs were separately trained for individual singer pairs determined in a round-robin fashion within intra-gender singers. The number of mixture components for the mel-cepstral coefficients was 128. The number of subjects was 6. We denote the spectral feature extracted based on STRAIGHT analysis as STRAIGHT mel-cepstrum and the spectral feature extracted simple spectral feature extraction as FFT mel-cepstrum.

5.4.2. Objective evaluation

For the objective evaluation, at first, we compared mel-cepstrum distortions (Mel-CD) to evaluate the effectiveness of the parameter transformation technique and frame-based GV post-filter for the low delay conversion. The Mel-CDs were calculated in the same manner as Equation 3.4.1 described in Section 3.4. Maximum likelihood parameter generation was hired as a parameter generation technique in this evaluation. We compared several techniques as follows:

DIFFSVC (w/o GV)

DIFFSVC without considering GV using STRAIGHT mel-cepstrum as an input feature,

Table 5.1.: Mel-cepstral distortions of low-delay statistical waveform modification techniques.

	Mel-CD [dB]
Method	Target singer
DIFFSVC (w/o GV)	4.54
RT-DIFFSVC (w/o GVPF)	4.86
RT-DIFFSVC (w/ GVPF, w/ intra)	5.03
RT-DIFFSVC (w/ GVPF, w/o intra)	5.19

RT-DIFFSVC (w/o GV)

DIFFSVC based on the real-time DIFFGMM without considering GV using FFT mel-cepstrum as an input feature vector,

RT-DIFFSVC (w/ GVPF w/ intra)

DIFFSVC based on the real-time DIFFGMM with frame-based GV post-filter using FFT mel-cepstrum as an input feature vector and intra-singer SVC to calculate the GV of converted feature trajectory,

RT-DIFFSVC (w/ GV, w/o intra)

DIFFSVC based on the real-time DIFFGMM with frame-based GV post-filter using FFT mel-cepstrum as an input feature vector.

Table 5.1 indicates that the Mel-CDs of several conversion techniques. It can be said that there is little quality degradation due to parameter transformation technique for low-delay conversion because the difference of Mel-CDs between “DIFFSVC (w/o GV)” and “RT-DIFFSVC (w/o GV)” is small. Therefore, it is considered that there is small sound quality degradation due to the use of the FFT mel-cepstrum as an input feature vector. We can see that the Mel-CD tends to be bigger when considering frame-based GV post-filter. This tendency is also observable in the offline conversion described in 3.2. There is small difference between “RT-DIFFSVC (w/ GVPF, w/ intra)” and “RT-DIFFSVC (w/ GVPF, w/o intra)”. Therefore, it is expected that the sound qualities of the converted voice are not difference.

In the objective evaluation, we also evaluated the delay of the low-delay statistical waveform modification technique. We calculated the delay τ based on cross-

Table 5.2.: Conversion delay of low-delay statistical waveform modification.

F_0 transformation ratio	delay [ms]
0.5	76.62
1.0 (w/o F_0 transformation)	35.00
1.0 (w/ F_0 transformation)	60.00
2.0	62.72

correlation function as follows:

$$\tau = argmax_{\tau} \sum_{\tau=0}^{\frac{N}{2}-1} \sum_{n=0}^{\frac{N}{2}-1} s_o(n) s_i(n+\tau), \tag{5.8}$$

where $s_i(n)$ and $s_o(n)$ denotes a sample of the input and converted waveforms, respectively. N is the number of samples of a sentence. The number of delay components for the low-delay conversion technique was set to 3. Note that the input and converted waveforms were recorded at the same time.

Table 5.2 illustrates the result of delay in the low-delay statistical waveform modification. In the intra-gender conversion, it is not necessary to perform F_0 transformation. Therefore, the delay of the low-delay conversion is equivalent to the conventional low-delay conversion. When the F_0 transformation is enabled, the delay tends to be bigger even when the F_0 transformation set to 1.0. This is because the waveform similarity-based over-lap add (WSOLA) requires proceedings frame of the input waveform. Moreover, when the F_0 transformation ratio set to 0.5 or 2.0, the delay of conversion significantly increases. This delay is quite large, but it is considered that the user can avoid a bad effect of his/her auditory feedback by practice.

5.4.3. Subjective evaluation

In the subjective evaluation, we evaluated the effectiveness of the frame-based GV post-filter for the statistical waveform modification. Two preference tests were evaluated to compare singing VC based on direct waveform modification (DIFFSVC) without GV post-filter (“w/o GVPF”) and DIFFSVC with proposed frame-based GV post-filter (“w/ GVPF”). The first preference test evaluated sound quality of the converted singing voices of the “w/o GVPF” and “w/ GVPF”. The converted singing voice sam-

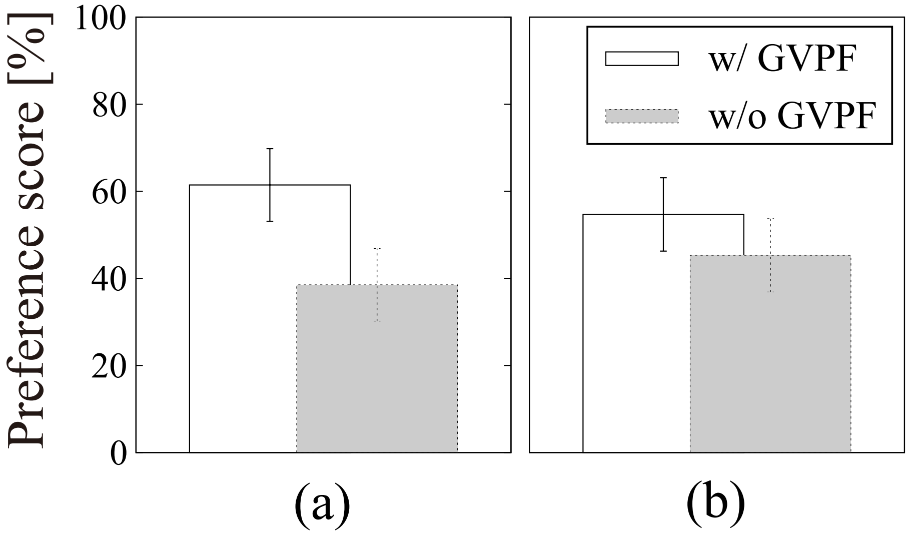


Figure 5.4.: Results of preference tests on (a) speech quality of converted singing voice and (b) conversion accuracy on singer individuality.

ples of the “w/ GVPF” and the “w/o GVPF” methods for the same phrase were presented to subjects in random order. The subjects selected which sample had better sound quality. The second preference test evaluated the conversion accuracy on singer identity of the converted singing voices. A natural singing voice sample of the target singer was presented to the subjects first as a reference. Then, the converted singing voice samples of the “w/ GVPF” and the “w/o GVPF” methods for the same phrase were presented in random order. The subjects selected which sample was more similar to the reference natural singing voice in terms of singer identity. The number of subjects in the first and second evaluation was 6 and each listener evaluated 32 sample pairs. All subjects don’t specialize in audio. Subjects were allowed to replay each sample pair as many times as necessary. Note that we have performed batch type conversion using maximum likelihood parameter generation [10] in the subjective evaluation because the conversion accuracy of low-delay conversion [25] and maximum likelihood parameter generation [10] are almost same.

Figure 5.4(a) indicates the results of preference test for the sound quality of the converted voice. We can see that “w/ GVPF” has higher sound quality compared with “w/o GVPF”. Figure 5.4(b) illustrates the results of conversion accuracy on singer individuality. It can be said that the conversion accuracies of “w/ GVPF” and “w/o

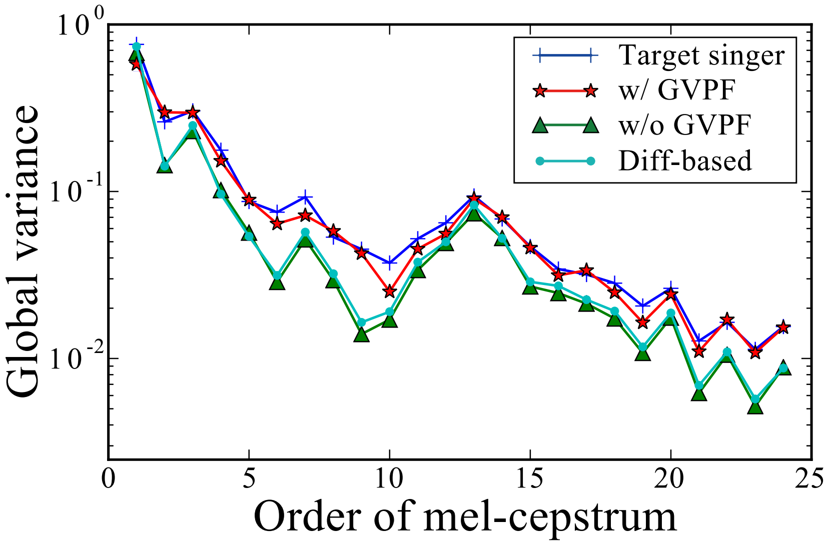


Figure 5.5.: GVs of mel-cepstral sequences of converted voices.

GVPF" are almost equal.

5.4.4. Analysis of converted feature trajectory

Figure 5.5 indicates the GVs of converted spectral feature extracted from converted singing voice. "Diff-based" indicates the GV of converted spectral feature using GV post-filtering against estimated spectral feature differential. We can see that the GV of "Diff-based" method can not restore the GV of the converted spectral features. On the other hand, the GV of "w/ GVPF" is restored compared with the GV of "w/o GVPF".

5.5. Summary

In this chapter, in order to implement real-time statistical waveform modification systems for voice conversion and voice timbre control, we have proposed following techniques: 1) parameter transformation technique for low-delay statistical waveform modification and 2) frame-based GV post-filtering for statistical waveform modification.

Section 5.2: This section has illustrated our implemented VC and voice timbre control systems.

Sections 5.3 and 5.4 : This section has proposed several techniques to implement the real-time conversion systems as follows: 1) parameter transformation technique for low-delay statistical waveform modification and 2) frame-based GV post-filtering for statistical waveform modification. The experimental results demonstrated that the frame-based GV post-filtering makes it possible to improve the sound quality of converted singing voice while maintaining conversion accuracy.

6. Conclusions

6.1. Summary of thesis

Voice conversion (VC) is a potential technique for enabling us to produce speech sounds beyond our own physical constraints. However, the VC framework has not yet been used in practice because the sound quality of the converted voice is significantly degraded compared with that of a natural speech waveform. One of the major factors causing the quality degradation is the waveform generation process using a vocoder. For the vocoding process, the sound quality degradation usually arises owing to various factors. To address this issue, in this thesis, we proposed VC and voice factor control techniques with statistical waveform modification.

In Chapter 2, we introduced traditional vocoding and VC frameworks based on Gaussian mixture models (GMM).

In Chapter 3, we addressed the sound quality degradation caused by vocoding in VC using speaking and singing voices. At first, in order to address the sound quality degradation of the converted singing voice, we proposed intra-gender singing VC (SVC) based on direct waveform modification using the spectral differential (DIFFSVC). Then, to make it possible to apply the intra-gender DIFFSVC to intra/inter-gender VC based on direct waveform modification using spectral differential (DIFFVC), we proposed several F_0 transformation techniques for the source voice. The experimental results confirmed that our proposed methods achieved higher sound quality of the converted voice than the conventional VC based on GMM using vocoding for waveform generation in both intra/inter-gender conversion.

In Chapter 4, we proposed a singing voice factor control technique based on voice timbre expression words while retaining singer individuality. We focused on the perceived age of the singing voice as a voice timbre expression word, because the perceived age is one of the most generic voice timbre expression words regardless of the user. First, we investigated several acoustic features of the singing voice affecting on the perceived age. Then, we proposed a voice timbre control technique based on the perceived age while retaining singer individuality. Moreover, to develop a better-controllable, higher quality, and more flexible voice timbre control framework, we also proposed the following techniques: 1) gender-dependent modeling, 2) voice timbre control based on direct waveform modification using a spectrum differential, and 3) an unsupervised adaptation method based on maximum a posteriori (MAP) estimation. The experimental results confirmed that our proposed voice timbre control methods make it possible to control voice timbre based on the perceived age while retaining singer identity.

In Chapter 5, we introduced our implemented real-time VC and voice timbre control systems based on statistical waveform modification. In order to implement real-time VC and voice timbre control systems, we proposed 1) a parameter transformation technique for the small-delay conversion algorithm and 2) a frame-based GV postfiltering for statistical waveform modification. The experimental results indicated that our proposed techniques make it possible to convert voice timbre with a sound quality comparable that obtained with the conventional technique in real time.

6.2. Future work

6.2.1. Voice conversion

In this thesis, although we struggled with the sound quality degradation of the converted voice in VC, an absolute VC system, which would enable us to completely convert the speaker/singer individuality without any sound degradation, has not yet been attained. In order to implement such a VC system, we consider that it would be meaningful to achieve the following.

Improvement of spectral conversion accuracy: In this thesis, we included a conditional probability density function of the GMM between a source and target speakers/singers for acoustic feature mapping. It is considered that more sophisticated acoustic modeling techniques such as discriminative training [74], Gaussian process regression [19,20], and deep neural networks [21,22,75], will reduce the sound quality degradation by incorporating statistical waveform modification. It is also important to focus on the factors such as error residual components of the converted feature trajectory, that are not converted in the statistical conversion.

F_0 **transformation without sound quality degradation:** For the inter-gender conversion, we proposed several F_0 transformation techniques for the statistical waveform modification. These F_0 transformation techniques usually cause slight sound quality degradation. In particular, the sound quality tends to be degraded when the F_0 transformation ratio is large. In order to improve the sound quality of the inter-gender conversion, it is necessary to improve the F_0 transformation technique.

Implementation of duration and time-variant F_0 transformation: In the GMM-based VC, duration and prosodic components are usually maintained after conversion. Although these restrictions to keep the prosodic components are meaningful in implementing a real-time VC system, to improve the conversion accuracy of speaker/singer individuality, it is worthwhile to implement a transformation technique of these components.

Direct modeling of waveform signals: WaveNet [23] can estimate not acoustic features such as F_0, aperiodicity, and spectral features, but waveform samples directly from linguistic features and F_0 in text-to-speech synthesis. Applying this technique to VC, significant improvements in sound quality are expected to be achieved.

When the absolute VC system has been established, the boundary between speakers and singers is assumed to vanish. Although this may be accepted by users who want to expand their speaking/singing expression, those who have superb abilities of speaking/singing may not be pleased with the VC system because it may also erase such abilities. Also, the VC system may possibly be used in crimes such as remittance fraud. We must not only focus on the effectiveness of the VC system, but also continue to consider and deal with problems related to VC.

6.2.2. Voice timbre control

For voice timbre control, although we only focus on the perceived age as a voice timbre control cue, the proposed voice timbre control technique is capable of controlling other voice timbre expression words. In related works, attempts to convert voice timbre based on the other voice timbre expression words [16, 76] have been made. Those groups reported that it is possible to control the voice timbre of a source speaker using other voice timbre expression words similarly to using the perceived age.

In the perceived age control, we did not perform any prosodic feature modification because the singer individuality significantly depends on the prosodic features compared with the segmental features. Therefore, if the prosodic features are controlled, the singer individuality will be lost. On the other hand, if the singer individuality need not be retained, it is worthwhile to manipulate the prosodic features using prosody control techniques [77, 78].

6.2.3. Real-time VC systems

In this thesis, we implemented real-time VC and voice timbre control systems with statistical waveform modification. We confirmed that our proposed systems greatly improve the sound quality of the converted voice compared with the conventional small-delay conversion system in any situation. This improvement mainly results from avoiding the F_0 extraction process, because the accuracy of F_0 extraction depends on the sound environment. However, there still remains some problems for practical use. First, the delay in real-time conversion tends to be very large when the F_0 transformation technique is used. This delay makes speaking to the system difficult via our auditory feed-back mechanism, because delayed auditory feed-back may introduce stammering [79]. In order to improve the conversion accuracy for speaker/singer individuality, it is necessary to decrease the delay to enable the auditory feed-back mechanism. Next, the conversion accuracy of the converted voice depends on the pair of source and target speakers. To use the VC system, stability of the conversion accuracy is important. Therefore, it is essential to improve the conversion accuracy of not only off-line conversion but also the real-time VC system.

A. The NU-NAIST VC system in VCC 2016

This chapter presents the NU-NAIST voice conversion (VC) system for the Voice Conversion Challenge 2016 (VCC 2016) developed by a joint team of Nagoya University (NU) and Nara Institute of Science and Technology (NAIST). VC research has been continued from about 20 years ago as a technique to convert speaker identity of a source speaker into that of a target speaker. However, it is difficult to directly compare their performances because there is no unified evaluation framework. In order to compare various VC techniques on identical training and evaluation speech data [80], VCC 2016 was held [52]. In this appendix, we describe additional results of the NU-NAIST VC system in VCC 2016.

A.1. Experimental evaluation

In this section, we describe results of the VCC 2016 to demonstrate performance of the NU-NAIST VC system. Moreover, we compare the following three systems:

- DIFFVC (EC): The NU-NAIST VC system submitted to the VCC 2016,
- VC: Our conventional VC system [13],
- DIFFVC: The NU-NAIST VC system w/o excitation conversion.

Note that the DIFFVC (EC) almost equals DIFFVC w/ STRAIGHT described in Section 3.4.2.

A.1.1. Experimental conditions

We evaluated speech quality and speaker identity of the converted voices to compare performance of the different VC systems in both intra-gender and cross-gender conversion tasks. We used the English speech database used in the VCC 2016. The number of source speakers was 5 including 3 females and 2 males, and that of the target speakers was 5 including 2 females and 3 males who were different from the source female and male speakers. The number of sentences uttered by each speaker was 216. The sampling frequency was set to 16 kHz.

STRAIGHT [12] was used to extract spectral envelopes, which was parameterized into the 1-24th mel-cepstral coefficients as the spectral feature. The frame shift was 5 ms. The mel log spectrum approximation (MLSA) filter [53] was used as the synthesis filter. As the source excitation features, we used F_0 and aperiodic components extracted with STRAIGHT [49]. The aperiodic components were averaged over five frequency bands, i.e., 0-1, 1-2, 2-4, 4-6, and 6-8 kHz, to be modeled with the GMM.

We used 162 sentences for training and the remaining 54 sentences were used for evaluation. The speaker-dependent GMMs were separately trained for all combinations of source and target speaker pairs. The number of mixture components for the mel-cepstral coefficients was 128 and for the aperiodic components was 64.

Two preference tests were conducted. In the first test, speech quality of the converted voices was evaluated. The converted voice samples generated by two different VC systems for the same sentences were presented to subjects in random order. The subjects selected which sample had better speech quality. In the second test, conversion accuracy in speaker identity was evaluated. A natural voice sample of the target speaker was presented to the subjects first as a reference. Then, the converted voice samples generated by two different VC systems for the same sentences were presented in random order. The subjects selected which sample was more similar to the reference natural voice in terms of speaker identity. The number of subjects was 10 and each listener evaluated 54 sample pairs in each evaluation. They were allowed to replay each sample pair as many times as necessary.

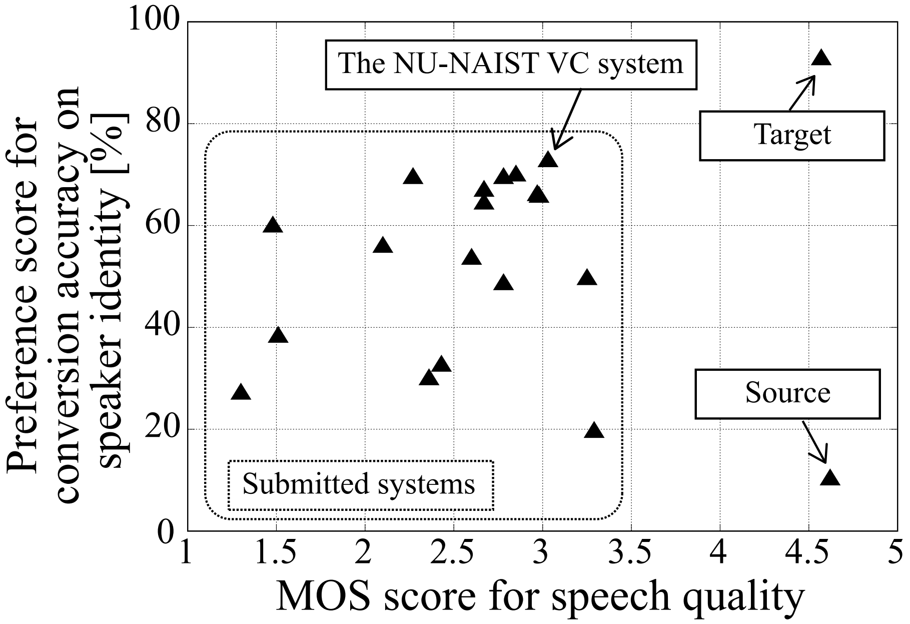


Figure A.1.: Sound quality and conversion accuracy on speaker identity in VCC 2016.

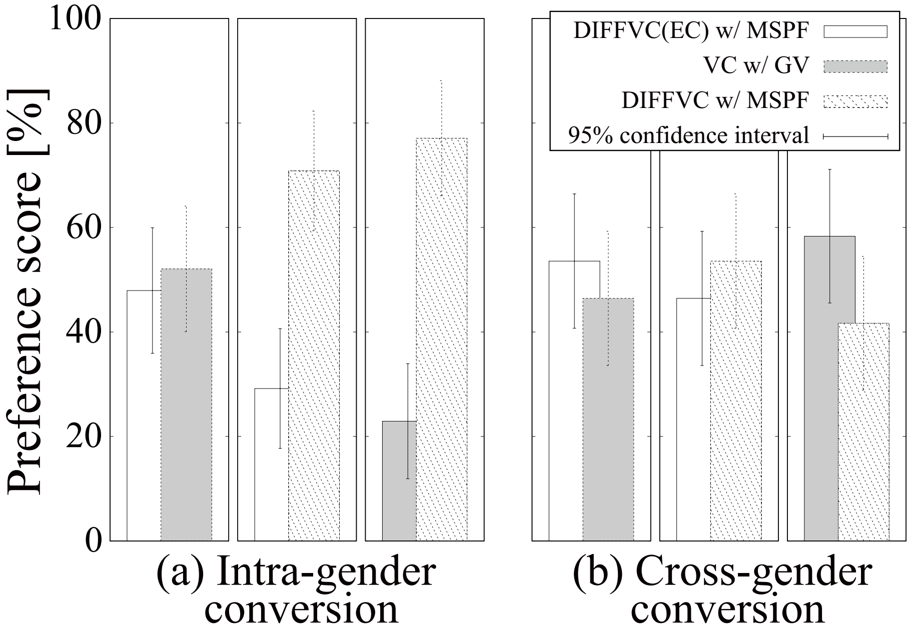


Figure A.2.: AB preference test for speech quality.

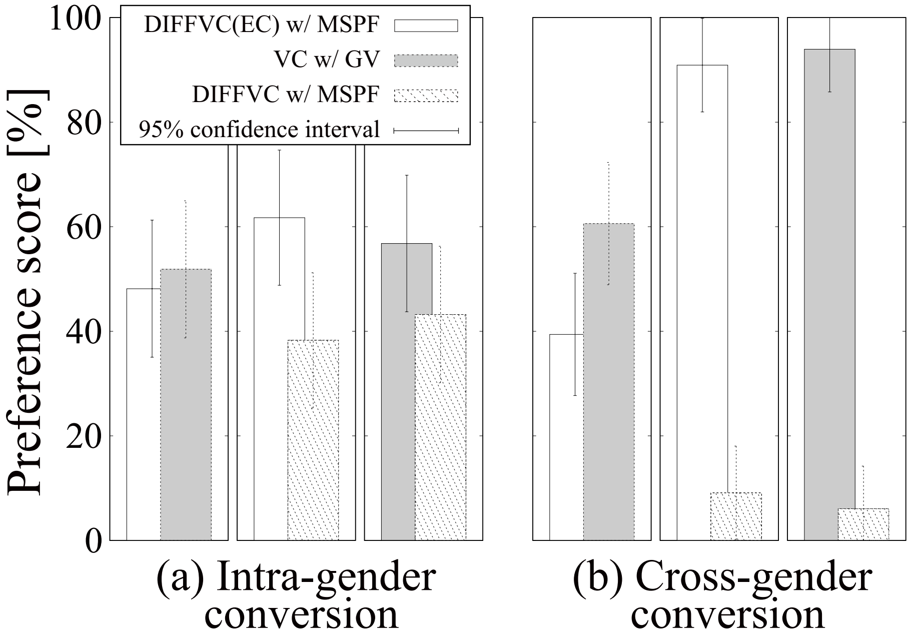


Figure A.3.: XAB test for conversion accuracy on speaker identity.

A.1.2. Results of the VCC 2016

Figure A.1 indicates an overall result of the VCC 2016. The NU-NAIST VC system achieved quite high speech quality over 3.0 of MOS and the best conversion accuracy (about 74%) among all submitted VC systems. In terms of the conversion accuracy, our system achieved successful performance even though very simple prosodic conversion was performed. However, it is observed that there is still a large gap between the converted voices and the natural target voices. It is expected that further improvements will be yielded by implementing a conversion method of prosodic patterns or asking the source speakers to mimic target prosodic patterns, which would be possible in several practical applications. In terms of speech quality, the NU-NAIST VC system causes serious quality degradation compared to natural voices, i.e., from 4.6 to 3.0 in MOS. This quality degradation is mainly caused by using a vocoder to perform the excitation conversion as shown in the next section. Therefore, it is expected that the converted speech quality will be significantly improved by developing a better analysis/synthesis technique than STRAIGHT.

A.1.3. Results of subjective evaluation

Figures A.2 (a) and (b) indicate the results of the preference test for speech quality. DIFFVC (EC) achieves equivalent speech quality compared to VC in both intra/cross-gender conversions. On the other hand, DIFFVC achieves significantly higher speech quality compared to the other two methods in the intra-gender conversion. This is because DIFFVC can avoid using vocoding to generate converted speech waveforms, making the conversion process free from various errors, such as F_0 extraction errors and unvoiced/voiced decision errors. Note that DIFFVC in cross-gender conversion condition does not result in any significant quality improvements as it suffers from mismatches between spectral envelope and F_0 in the cross-gender conversion.

Figures A.3 (a) and (b) indicate the results of the preference test for speaker identity. Although DIFFVC (EC) has equivalent conversion accuracy compared to VC in the intra-gender conversion, it tends to be degraded in the cross-gender conversion. It is expected that the residual spectral envelope preserved in the direct waveform modification process still includes speaker-dependent or gender-dependent features, and that this causes adverse effects on conversion accuracy.

These results suggest that 1) the NU-NAIST VC system demonstrating the best conversion accuracy and high speech quality in the VCC 2016 has an almost equivalent performance compared to the conventional VC system in both intra-gender and cross-gender conversions, and 2) the direct waveform modification technique achieves significantly higher converted speech quality compared to the conventional VC system if the excitation conversion is not necessary as in the intra-gender conversion, and therefore, there is still large room to improve the converted speech quality of the NU-NAIST VC system.

A.2. Summary

This chapter describes the NU-NAIST voice conversion (VC) system for the Voice Conversion Challenge 2016 (VCC 2016) developed by a joint team of Nagoya University (NU) and Nara Institute of Science and Technology (NAIST). In order to improve the quality of statistical VC based on Gaussian Mixture Model (GMM), we applied the following techniques: 1) voice conversion with direct waveform modification with spectral differential (DIFFVC), 2) speech parameter trajectory smoothing, 3) post-filtering based on modulation spectrum for DIFFVC, and 4) preprocessing for excitation conversion with F_0 and aperiodic component transformations using high-quality vocoding. The experimental results demonstrated that the NU-NAIST VC system was highly ranked in the VCC 2016, its performance was comparable to our conventional VC system, and the DIFFVC technique showed large potential to significantly improve the converted speech quality of the NU-NAIST VC system.

Acknowledgements

本学情報科学研究科の 中村 哲 教授には，主指導教官として，研究指導のみならず，様々な事柄をご指導頂きました．とりわけ，研究と社会の関係性や博士取得後のキャリアの展望など，一人の研究者として，活躍するための多くの示唆を頂きました．教えて頂いた事を深く心に刻み，卒業後も研究者として奮励したいと思います．心より深く感謝致します.

本学情報科学研究科の 杉本 謙二 教授には，副指導教官として，本論文をより精錬する上で多くのご指導をいただきました．心より感謝致します.

名古屋大学の 戸田 智基 教授（前 奈良先端科学技術大学院大学 准教授）には，副指導教官として，研究全般における数々の有益な御助言や熱心な御指導をはじめとし，博士課程を支え続けて頂きました．また，原稿，資料作成時には，大変丁寧な御指導を頂きました．心より厚く感謝致します.

本学情報科学研究科の Sakriani Sakti 助教 には，研究に関して多くのご助言を頂きました．また，研究以外の多くの相談にも丁寧に答えて頂きました．心より感謝致します.

産業技術総合研究所の 後藤 真孝 博士および 中野 倫靖 博士には，博士前期課程一年時に参加させて頂いた研究実習より博士後期課程取得までご指導頂きました．産総研での夏季実習では，多くの学びやかけがえのない友人を得ることが出来ました．得られた繋がりを大切に今後も励んでいきたいと思います．深謝致します.

Carnegie Mellon University の Graham Neubig 助教（前 奈良先端科学技術大学院大学 助教）には，博士前期課程時を中心に，貴重な御助言を頂きました. 心より感謝致します.

株式会社ドワンゴの 土井 啓成 博士には，研究開始時に非常に熱心な御指導をはじめとし，研究生活との向き合い方を背中で教えて頂きました．感謝致します.

知能コミュニケーション研究室の諸氏には，主観評価実験の協力，数多くの有益な助言をいただきました．また，Local Optimum や AHC 肉体改造部を始めとした素晴らしい日々を共に過ごすことが出来ました．これらの活動は研究成果に

は直結しないものの，有益な経験だったと思います．心より感謝致します.

知能コミュニケーション研究室の 松田真奈美 秘書には，庶務の手続きのみならず，研究生活において多くの楽しみや喜び，話題を提供して頂きました．心より感謝致します.

博士後期課程において，フランス留学時の研究指導を引き受けて頂いた GIPSA-LAB の Gérard Bailly 博士および訪問講演をさせて頂いた University of Edinburgh の Simon King 教授，名古屋工業大学の 徳田 恵一 教授には，研究へのご助言を頂きました．感謝致します.

クリムゾンテクノロジー株式会社の飛河 和生 様および高橋 賢一 様には共同研究を進める上で多くの助言やサポートを頂きました．また，研究成果を技術に変える事の素晴らしさを体験させて頂く事が出来ました．感謝致します.

本研究の一部は，日本学術振興会 特別研究員（DC2）の特別研究員奨励費16J10726により実施されました．研究費の助成により博士後期課程の研究を一層飛躍させる事が出来ました．感謝致します.

最後に，学業生活を陰ながら支えて頂いた家族に感謝致します.

Publication List

Journal papers

1. Kazuhiro Kobayashi, Tomoki Toda, Hironori Doi, Tomoyasu Nakano, Masataka Goto, Satoshi Nakamura. "Voice Timbre Control Based on Perceived Age in Singing Voice Conversion," IEICE Transactions on Information and Systems, Vol. E97-D, No. 6, pp. 1419-1428, June 2014.

2. Kazuhiro Kobayashi, Tomoki Toda, Tomoyasu Nakano, Masataka Goto, Satoshi Nakamura. "Improvements of voice timbre control based on perceived age in singing voice conversion,”IEICE Transactions on Information and Systems, Vol. E99-D, No. 11, pp. 2767-2777, Nov. 2016.

International conferences

1. Kazuhiro Kobayashi, Tomoki Toda and Satoshi Nakamura, "F0 transformation techniques for statistical voice conversion with direct waveform modification with spectral differential,”Proc. SLT, pp. 693-700, San Diego, USA, Dec. 2016.

2. Kazuhiro Kobayashi, Shinnosuke Takamichi, Tomoki Toda and Satoshi Nakamura, "The NU-NAIST voice conversion system for the Voice Conversion Challenge 2016, ” Proc. INTERSPEECH, pp. 1667-1671, San Francisco, USA, Sep. 2016.

3. Kazuhiro Kobayashi, Tomoki Toda and Satoshi Nakamura, "Implementation of F0 transformation for statistical singing voice conversion based on direct waveform modification,”Proc. ICASSP, pp. 5670-5674, Shanghai, China, Mar. 2016.

4. Kazuhiro Kobayashi, Tomoki Toda, Graham Neubig, Sakriani Sakti and Satoshi Nakamura, “Statistical singing voice conversion based on direct waveform modification with global variance," Proc. INTERSPEECH, pp. 2754–2758, Dresden , Sept. 2015.

5. Kazuhiro Kobayashi, Tomoki Toda, Tomoyasu Nakano, Masataka Goto, Graham Neubig, Sakriani Sakti and Satoshi Nakamura, "Gender-dependent spectrum dif-

ferential models for perceived age control based on direct waveform modification in singing voice conversion," Proc. APSIPA ASC, Cambodia, Dec. 2014.

6. Kazuhiro Kobayashi, Tomoki Toda, Graham Neubig, Sakriani Sakti and Satoshi Nakamura, "Statistical singing voice conversion with direct waveform modification based on the spectrum differential," Proc. INTERSPEECH, pp. 2514-2518, Singapore, Sept. 2014.

7. Kazuhiro Kobayashi, Tomoki Toda, Tomoyasu Nakano, Masataka Goto, Graham Neubig, Sakriani Sakti and Satoshi Nakamura, "Regression approaches to perceptual age control in singing voice conversion," Proc. ICASSP, pp. 7954-7958, Florence, Italy, May. 2014.

8. Kazuhiro Kobayashi, Hironori Doi, Tomoki Toda, Tomoyasu Nakano, Masataka Goto, Graham Neubig, Sakriani Sakti, Satoshi Nakamura, “ An investigation of acoustic features for singing voice conversion based on perceptual age," Proc. INTERSPEECH, pp. 1057-1061, Lyon, France, Aug. 2013.

Technical reports

1. 小林 和弘, 戸田 智基, 中村 哲, "差分スペクトル補正に基づく歌声声質変換のための F0 変換の評価 ”, 信学技報, Vol. 115, No. 523, SP2015-112, pp. 105-110, Mar. 2016.

2. 小林 和弘, 戸田 智基, 中村 哲, "差分スペクトル補正による統計的歌声声質変換とパラメータ生成法 ”, 信学技報, Vol. 115, No. 253, SP2015-60, pp.7-12, Oct. 2015.

3. 小林 和弘, 戸田 智基, 中野 倫靖, 後藤 真孝, Graham Neubig, Sakriani Sakti, 中村 哲, "知覚年齢をリアルタイムに制御可能な歌声声質制御インタフェース ”, 第22回インタラクティブシステムとソフトウェアに関するワークショップ（WISS 2014）論文集, デモ, 3B-07, Nov. 2014.

4. 小林和弘, 戸田智基, 中野倫靖, 後藤真孝, ニュービッグ グラム, サクリアニ サクティ, 中村 哲, "統計的手法に基づく歌声の知覚年齢制御法", 信学技報, Vol. 114, No. 52, SP2014-30, pp. 321-326, May 2014.

5. 小林和弘, "統計的手法に基づく歌声声質変換 ”, 第 100 回 情報処理学会 音声言語情報処理研究会, Jan. 2014.

6. 小林 和弘, 戸田 智基, Graham Neubig, Sakriani Sakti, 中村 哲, "統計的歌声声質変換における知覚年齢に基づく声質制御 ”, 信学技報, Vol. 113, No. 308, SP2013-71, pp. 1-6, Nov. 2013.

7. 小林 和弘, 土井 啓成, 戸田 智基, 中野 倫靖, 後藤 真孝, Graham Neubig, Sakriani Sakti, 中村 哲, "知覚年齢に沿った歌声声質制御のための音響特徴量の調査", 情報処理学会研究報告, Vol.2013-MUS-99 No.44, pp. 1-6, 2013.

Domestic conferences

1. 小林 和弘, 戸田 智基, 中村 哲, "差分スペクトル補正に基づく声質変換における F0 変換法の調査 ”, 日本音響学会春季研究発表会, 1-6-11, pp. 229-230, Mar. 2017.

2. Kazuhiro Kobayashi, Tomoki Toda, Satoshi Nakamura, "Low delay statistical singing voice conversion with direct waveform modification based on spectral differential considering global variance, ” 5th Joint Meeting of the ASA and the ASJ, 1aSC16, Hawaii, USA, Nov. 2016.

3. 小林 和弘, 戸田 智基, 中村 哲, "系列内変動を考慮した差分スペクトル補正に基づく短遅延歌声声質変換 ”, 日本音響学会秋季研究発表会, 1-R-39, pp. 337-338, Mar. 2016.

4. 小林 和弘, 戸田 智基, Graham Neubig, Sakriani Sakti, 中村 哲, "差分スペクトル補正による歌声声質変換のための F0 変換に関する検討 ”, 日本音響学会秋季研究発表会, 3-1-10, pp. 249-250, Sep. 2015.

5. 小林 和弘, 戸田 智基, Graham Neubig, Sakriani Sakti, 中村 哲: "差分スペクトル補正に基づく歌声声質変換におけるパラメータ生成法に関する調査 ”, 日本音響学会春季研究発表会, 3-2-6, Mar. 2015.

6. 小林 和弘, 戸田 智基, 中野 倫靖, 後藤 真孝, ニュービッグ グラム, サクリアニ サクティ, 中村 哲, "性別依存重回帰混合正規分布モデル基づく差分スペクト

ル補正による歌声の知覚年齢制御法", 日本音響学会秋季研究発表会, 3-7-4, Sept. 2014.

7. 小林 和弘, 戸田 智基, 中野 倫靖, 後藤 真孝, G. Neubig , S. Sakuti, 中村 哲, " 歌唱音声の統計的知覚年齢制御", OngaCREST シンポジウム, Aug. 2014.

8. 小林 和弘, 戸田 智基, Graham Neubig, Sakriani Sakti, 中村 哲: "差分スペクトル補正に基づく歌声声質変換 ", 日本音響学会春季研究発表会, 3-6-4, Mar. 2014.

9. 小林 和弘, 戸田 智基, Graham Neubig, Sakriani Sakti, 中村 哲: "統計的歌声声質変換における知覚年齢に沿った声質制御 ", 日本音響学会秋季研究発表会, 3-7-8, pp. 1479-1480, Sep. 2013.

10. 小林 和弘, 土井 啓成, 戸田 智基, 中野 倫靖, 後藤 真孝, Graham Neubig, Sakriani Sakti, 中村 哲, "歌声の知覚年齢に沿った声質制御に向けた音響特徴量の調査 ", 日本音響学会春季研究発表会, 2-7-14, pp. 347-348, Mar. 2013.

Awards

1. 奈良先端科学技術大学院大学 最優秀学生賞（博士）

2. 奈良先端科学技術大学院大学支援財団 NAIST 最優秀学生賞（博士）

3. 奈良先端科学技術大学院大学 優秀学生奨励賞

4. 日本音響学会関西支部 最優秀研究奨励賞

5. 日本音響学会 第 10 回（2014 年秋季研究発表会） 学生優秀発表賞

Research talks

1. Kazuhiro Kobayashi, "Statistical voice conversion based on direct waveform modification with spectral differential," 名古屋工業大学, Nov. 18th, 2016. (Invited talk)

2. Kazuhiro Kobayashi, "Statistical singing voice conversion based on direct waveform modification with spectral differential," University of Edinburgh, Feb. 11th, 2016. (Visiting talk)

3. Kazuhiro Kobayashi, "Statistical singing voice conversion based on direct waveform modification with spectral differential," GIPSA-lab, Feb. 3th, 2016. (Visiting talk)

Related publications

International conferences

1. Soichi Yamane, Kazuhiro Kobayashi, Tomoki Toda, Tomoyasu Nakano, Masataka Goto, Satoshi. Nakamura. "An estimation method of voice timbre evaluation values using feature extraction with Gaussian mixture model based on reference singer, ” Proc. ICASSP, pp. 5265-5269, Shanghai, China, Mar. 2016.

2. Shinnosuke Takamichi, Kazuhiro Kobayashi, Kou Tanaka, Tomoki Toda, Satoshi Nakamura, "The NAIST text-to-speech system for the Blizzard Challenge 2015, ” Proc. Blizzard Challenge Workshop, 4 pages, Berlin, Germany, Sep. 2015.

3. Patrick Lumban Tobing, Kazuhiro Kobayashi, Tomoki. Toda Graham Neubig, Sakriani Sakti, Satoshi Nakamura, "Articulatory controllable speech modification based on Gaussian mixture models with direct waveform modification using spectrum differential, ” Proc. INTERSPEECH, pp. 3350-3354, Dresden, Germany, Sep. 2015.

Technical reports

1. 山根 壮一, 小林 和弘, 戸田 智基, 中野 倫靖, 後藤 真孝, Graham Neubig, Sakriani Sakti, 中村 哲, "歌声合成システムの音源データ検索のための声質評価値推定 ”, 情報処理研報, Vol. 2015-MUS-108, No. 6, pp. 1-6, Aug. 2015.

2. 久保 和隆, 小林 和弘, 戸田 智基, Graham Neubig, Sakriani Sakti, 中村 哲, "重回帰混合正規分布モデルに基づく声質制御における制御パラメータの設計 ”, 信学技報, Vol. 114, No. 303, SP2014-101, pp. 65-70, Nov. 2014.

Domestic conferences

1. 武山知弘, 小林和弘, 田尻祐介, 戸田智基, 武田一哉"統計的音声波形変換に基づく雑音環境下における音声了解度向上に向けた調査, ” 音講論, 1-6-12, pp. 231-232, Mar. 2017.

2. 山根 壮一, 小林 和弘, 戸田 智基, 中野 倫靖, 後藤 真孝, 中村 哲, "歌声合成システムの音源データに対する声質評価値に基づく声質制御, ” 音講論, 2-2-8, pp. 247-248, Mar. 2016.

3. Patrick Lumban Tobing, Kazuhiro Kobayashi, Tomoki Toda, Graham Neubig, Sakriani Sakti, S. Nakamura, "An evaluation of articulatory controllable speech modification based on Gaussian mixture models with direct waveform modification, ” 音講論, 2-1-3, pp. 221-222, Sep. 2015.

4. 山根 壮一, 小林 和弘, 戸田 智基, 中野 倫靖, 後藤 真孝, Graham Neubig, Sakriani Sakti, 中村 哲, "歌声合成による学習データ生成を利用した歌声の声質評価値推定法, ” 音講論, 3-1-9, pp. 247-248, Sep. 2015.

5. 久保 和隆, 小林 和弘, 戸田 智基, Graham Neubig, Sakriani Sakti, 中村 哲, " 重回帰混合正規分布モデルに基づく声質制御における精度改善 ” , 音講論, 2-2-7, pp. 265-266, Mar. 2015.

6. Patrick Lumban Tobing, Kazuhiro Kobayashi, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura, "Articulatory controllable speech modification based on gaussian mixture models with direct waveform modification using spectrum differential, ” 音講論, 2-2-8, pp. 267-268, Mar. 2015.

Awards

1. 第11回日本音響学会学生優秀発表賞（受賞者：Patrick Lumban Tobing）

References

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, “Voice conversion through vector quantization,” *J. Acoust. Soc. Jpn (E)*, vol. 11, no. 2, pp. 71–76, 1990.

[2] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, “A hybrid approach to electrolaryngeal speech enhancement based on noise reduction and statistical excitation generation,” *IEICE Trans. Inf. and Syst.*, vol. E97-D, no. 6, pp. 1429–1437, June 2014.

[3] T. Toda, A. Black, and K. Tokuda, “Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model,” *Speech Communication*, pp. 215–227, Mar. 2008.

[4] P. Tobing, T. Toda, G. Neubig, S. Sakti, S. Nakamura, and A. Purwarianti, “Articulatory controllable speech modification based on statistical feature mapping with gaussian mixture model,” *Proc. INTERSPEECH*, pp. 2298–2302, Sept. 2014.

[5] A. Barbulescu, T. Hueber, G. Bailly, and R. Ronfard, “Audio-visual speaker conversion using prosody features,” *Proc. AVSP*, pp. 11–16, Aug. 2013.

[6] Y. Tajiri, K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura., “Non-audible murmur enhancement based on statistical conversion using air- and body-conductive microphones in noisy environments,” *Proc. INTERSPEECH*, pp. 2769–2773, Sep. 2015.

[7] T. Toda, “Augmented speech production based on real-time statistical voice conversion,” *Proc. GlobalSIP*, pp. 755–759, Dec. 2014.

[8] Y. Stylianou, O. Cappé, and E. Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE Trans. SAP*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[9] Y. Kawakami, H. Banno, and F. Itakura, “GMM voice conversion of singing voice using vocal tract area function,” *IEICE technical report. Speech (Japanese edition)*, vol. 110, no. 297, pp. 71–76, Nov. 2010.

[10] T. Toda, A. W. Black, and K. Tokuda, “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” *IEEE Trans. ASLP*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.

[11] S. Takamichi, T. Toda, A. Black, G. Neubig, S. Sakti, and S. Nakamura, “Postfilters to modify the modulation spectrum for statistical parametric speech synthesis,” *IEEE Trans. ASLP*, vol. 24, no. 4, pp. 755–767, Jan. 2016.

[12] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f_0 extraction: Possible role of a repetitive structure in sounds,” *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, Apr. 1999.

[13] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, “Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation,” *Proc. INTERSPEECH*, pp. 2266–2269, Sept. 2006.

[14] ——, “Many-to-many eigenvoice conversion with reference voice,” *Proc. INTERSPEECH*, pp. 1623–1626, Sept. 2009.

[15] H. Doi, T. Toda, T. Nakano, M. Goto, and S. Nakamura, “Singing voice conversion method based on many-to-many eigenvoice conversion and training data generation using a singing-to-singing synthesis system,” *Proc. APSIPA ASC*, Nov. 2012.

[16] K. Ohta, T. Toda, Y. Ohtani, H. Saruwatari, and K. Shikano, “Adaptive voice-quality control based on one-to-many eigenvoice conversion,” *Proc. INTERSPEECH*, pp. 2158–2161, Sept. 2010.

[17] H. Kido and H. Kasuya, “Everyday expressions associated with voice quality normal utterance extraction by perceptual evaluation,” *The Acoustical Society of Japan (Japanese edition)*, pp. 337–344, May 2001.

[18] A. Kanato, T. Nakano, M. Goto, and H. Kikuchi, “An automatic singing impression estimation method using factor analysis and multiple regression,” *Proc. ICMC SMC*, pp. 1244–1251, Sep. 2014.

[19] N. Pilkington, H. Zen, and M. Gales, “Gaussian process experts for voice conversion,” *Proc. INTERSPEECH*, pp. 2761–2764, Aug. 2011.

[20] N. Xu, Y. Tang, J. Bao, A. Jiang, X. Liu, and Z. Yang, “Voice conversion based on Gaussian processes by coherent and asymmetric training with limited training data,” *Speech Communication*, vol. 58, pp. 124–138, Mar. 2014.

[21] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, “Voice conversion using deep neural networks with layer-wise generative training,” *IEEE/ACM Trans. ASLP*, vol. 22, no. 12, pp. 1859–1872, Dec. 2014.

[22] L. Sun, S. Kang, K. Li, and H. Meng, “Voice conversion using deep bidirectional long short-term memory based recurrent neural networks,” *Proc. ICASSP*, pp. 4869–4873, Apr. 2015.

[23] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” *CoRR*, vol. abs/1609.03499, 2016.

[24] T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, “Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” *Proc. INTERSPEECH*, pp. 1076–1079, Sept. 2008.

[25] T. Toda, T. Muramatsu, and H. Banno, “Implementation of computationally efficient real-time voice conversion,” *Proc. INTERSPEECH*, Sept. 2012.

[26] H. Dudley, “Remaking speech,” *JASA*, vol. 11, no. 2, pp. 169–177, 1939.

[27] Y. Stylianou, “Applying the harmonic plus noise model in concatenative speech synthesis,” *IEEE Trans. SAP*, vol. 9, no. 1, pp. 21–29, 2001.

[28] D. Erro, I. Sainz, E. Navas, and I. Hernaez, “Harmonics plus noise model based vocoder for statistical parametric speech synthesis,” *IEEE J-STSP*, vol. 8, no. 2, pp. 184–194, 2014.

[29] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE Trans. Inf. and Syst.*, vol. E99-D, no. 7, pp. 1877–1884, 2016.

[30] W. Ryan and K. Burk, “Perceptual and acoustic correlates of aging in the speech of males,” *Journal of Communication Disorders*, vol. 7, no. 2, pp. 181–192, 1974.

[31] H. Kasuya, H. Yoshida, S. Ebihara, and H. Mori, “Longitudinal changes of selected voice source parameters,” *Proc. INTERSPEECH*, pp. 2570–2573, Sept. 2010.

[32] J. D. Harnsberger, W. S. Brown Jr., R. Shrivastav, and H. Rothman, “Noise and tremor in the perception of vocal aging in males,” *Journal of Voice*, vol. 24, no. 5, pp. 523–530, 2010.

[33] N. Minematsu, M. Sekiguchi, and K. Hirose, “Automatic estimation of one’s age with his/her speech based upon acoustic modeling techniques of speakers,” *Proc. ICASSP*, pp. 137–140, May. 2002.

[34] Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu, “Voice conversion with smoothed GMM and MAP adaptation,” *Proc. INTESPEECH*, pp. 1–4, Sept. 2003.

[35] C.-H. Lee and C.-H. Wu, “Map-based adaptation for speech conversion using adaptation data selection and non-parallel training,” *Proc. INTERSPEECH, 2006*, pp. 17–21, Sept. 2006.

[36] D. Erro, A. Moreno, and A. Bonafonte, “INCA algorithm for training voice conversion systems from nonparallel corpora,” *IEEE Trans. ASLP*, vol. 18, no. 5, pp. 944–953, 2010.

[37] “HMM-based speech synthesis system (HTS) http://hts.sp.nitech.ac.jp/.”

[38] H. Zen, K. Tokuda, and A. W. Black, “Statistical parametric speech synthesis,” *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[39] E. Moulines and F. Charpentier, “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones,” *Speech Communication*, vol. 9, no. 5, pp. 453–467, Dec. 1990.

[40] S. Imai, K. Sumita, and C. Furuichi, “Mel log spectrum approximation (MLSA) filter for speech synthesis,” *Electronics and Communications in Japan (Part I: Communications)*, vol. 66, no. 2, pp. 10–18, 1983.

[41] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” *Proc. ICASSP*, pp. 1315–1318, June 2000.

[42] S. Takamichi, T. Toda, A. Black, and S. Nakamura, “Modulation spectrum-based post-filter for GMM-based voice conversion,” *Proc. APSIPA ASC*, Dec. 2014.

[43] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, “Adaptive training for voice conversion based on eigenvoices,” *IEICE Trans. Inf. and Syst.*, vol. E93-D, no. 6, pp. 1589–1598, June 2010.

[44] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, “A compact model for speaker-adaptive training,” *Proc. ICSLP*, vol. 2, pp. 1137–1140, 1996.

[45] T. Anastasakos, J. McDonough, and J. Makhoul, “Speaker adaptive training: A maximum likelihood approach to speaker normalization,” *Proc. ICASSP*, vol. 2, pp. 1043–1046, 1997.

[46] D. T. Chappell and J. H. L. Hansen, “Speaker-specific pitch contour modeling and modification,” *Proc. ICASSP*, vol. 2, pp. 885–888, May 1998.

[47] M. Sambur, “Selection of acoustic features for speaker identification,” *IEEE Trans. ASSP*, vol. 23, no. 2, pp. 176–182, 1975.

[48] H. Zen, K. Tokuda, and T. Kitamura, “Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences,” *Computer Speech & Language*, vol. 21, no. 1, pp. 153–173, 2007.

[49] H. Kawahara, J. Estill, and O. Fujimura, “Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and system straight,” *Proc. MAVEBA*, Sept. 2001.

[50] W. Verhelst and M. Roelands, “An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech,” *Proc. ICASSP*, vol. 2, pp. 554–557, Apr. 1993.

[51] K. Kobayashi, T. Toda, and S. Nakamura, “Implementation of f0 transformation for statistical singing voice conversion based on direct waveform modification,” *Proc. ICASSP*, pp. 5670–5674, Mar. 2016.

[52] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, “The Voice Conversion Challenge 2016,” *Proc. INTERSPEECH*, pp. 1632–1636, Sept. 2016.

[53] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, “Mel-generalized cepstral analysis – a unified approach to speech spectral estimation,” *Proc. ICSLP*, pp. 1043–1045, 1994.

[54] J. Bonada, A. Loscos, and H. Kenmochi, “Sample-based singing voice synthesizer by spectral concatenation,” *Proc. SMAC*, pp. 1–4, 2003.

[55] H. Kenmochi and H. Ohshita, “VOCALOID – Commericial singing synthesizer based on sample concatenation,” *Proc. INTERSPEECH*, pp. 4011–4012, Aug. 2007.

[56] K. Saino, M. Tachibana, and H. Kenmochi, “A singing style modeling system for singing voice synthesizers.” *Proc. INTERSPEECH*, pp. 2894–2897, Sept. 2010.

[57] K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku, and K. Tokuda, “Recent development of the HMM-based singing voice synthesis system - Sinsy,” *Proc. SSW7*, pp. 211–216, Sept. 2010.

[58] M. Nishimura, K. Hashimoto, O. Keiichiro, Y. Nankaku, and K. Tokuda, “Singing voice synthesis based on deep neural networks,” *Proc. INTERSPEECH*, pp. 2478–2482, Sept. 2016.

[59] T. Nakano and M. Goto, “VocaListener: A singing-to-singing synthesis system based on iterative parameter estimation,” *Proc. SMC*, pp. 343–348, July 2009.

[60] T. Nakano and M. Goto, “Vocalistener2: A singing synthesis system able to mimic a user’s singing in terms of voice timbre changes as well as pitch and dynamics,” *Proc. ICASSP*, pp. 453–456, May 2011.

[61] M. Morise, M. Onishi, H. Kawahara, and H. Katayose, “v. morish’ 09: A morphing-based singing design interface for vocal melodies,” *Proc. ICEC*, pp. 185–190, Sept. 2009.

[62] F. Villavicencio and J. Bonada, “Applying voice conversion to concatenative singing-voice synthesis,” *Proc. INTERSPEECH*, pp. 2162–2165, Sept. 2010.

[63] T. Toda, Y. Ohtani, and K. Shikano, “One-to-many and many-to-one voice conversion based on eigenvoices,” *Proc. ICASSP*, pp. 1249–1252, Apr. 2007.

[64] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, “A style control technique for HMM-based expressive speech synthesis (speech and hearing),” *IEICE Trans. Inf. and Syst.*, vol. 90, no. 9, pp. 1406–1413, Sep. 2007.

[65] T. Nose and T. Kobayashi, “An intuitive style control technique in HMM-based expressive speech synthesis using subjective style intensity and multiple-regression global variance model,” *Speech Communication*, vol. 55, no. 2, pp. 347–357, 2013.

[66] M. Tachibana, T. Nose, J. Yamagishi, and T. Kobayashi, “A technique for controlling voice quality of synthetic speech using multiple regression HSMM,” *Proc. INTERSPEECH*, pp. 2438–2441, Sept. 2006.

[67] M. Goto and T. Nishimura, “AIST humming database: Music database for singing research,” *IPSJ SIG Notes (Technical Report) (Japanese edition)*, vol. 2005-MUS-61-2, pp. 7–12, Aug. 2005.

[68] W. Endres, W. Bambach, and G. Flösser, “Voice spectrograms as a function of age, voice disguise, and voice imitation,” *The Journal of the Acoustical Society of America*, vol. 49, no. 6B, pp. 1842–1848, 1971.

[69] S. E. Linville and J. Rens, “Vocal tract resonance analysis of aging voice using long-term average spectra,” *Journal of Voice*, vol. 15, no. 3, pp. 323–330, 2001.

[70] K. Kobayashi, T. Toda, D. H, T. Nakano, M. Goto, G.Neubig, S. Sakti, and S. Nakamura, “Voice timbre control based on perceived age in singing voice conversion,” *IEICE Trans. Inf. and Syst.*, vol. 97, no. 6, pp. 1419–1428, 2014.

[71] X. Li and J. Rekimoto, “Smartvoice: A presentation support system for overcoming the language barrier,” *Proc. CHI, ACM*, pp. 1563–1570, 2014.

[72] K. Yamamoto, “Possessing drums: An interface of musical instruments that assigns arbitrary timbres to personal belongings,” *Journal of Information Processing*, vol. 21, no. 2, pp. 274–282, Apr. 2013.

[73] K. Nakano, M. Morise, T. Nishiura, and Y. Yamashita, “Improvement of high-quality vocoder straight for vocal manipulation system based on fundamental frequency transcription,” *IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences (in Japanese)*, vol. E95-A, no. 7, pp. 563–572, July 2012.

[74] H. T. Hwang, Y. Tsao, H.-M. Wang, Y.-R. Wang, and S.-H. Chen, “Alleviating the over-smoothing problem in GMM-based voice conversion with discriminative training,” *Proc. INTERSPEECH*, p. 2013, Aug. 3062–3066.

[75] T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, “Voice conversion in high-order eigen space using deep belief nets,” *Proc. INTERSPEECH*, pp. 369–372, Aug. 2013.

[76] S. Yamane, K. Kobayashi, T. Toda, T. Nakano, M. Goto, and S. Nakamura, “An estimation method of voice timbre evaluation values using feature extraction with gaussian mixture model based on reference singer,” *Proc. ICASSP*, pp. 5265–5269, Mar. 2016.

[77] Y. Ohishi, D. Mochihashi, H. Kameoka, and K. Kashino, “Mixture of Gaussian process experts for predicting sung melodic contour with expressive dynamic fluctuations,” *Proc. ICASSP*, pp. 3742–3746, May 2014.

[78] Y. Ikemiya, K. Yoshii, and K. Itoyama, “Singing voice analysis and editing based on mutually dependent F0 estimation and source separation,” *Proc. ICASSP*, pp. 574–577, Apr. 2015.

[79] A. J. Yates, “Delayed auditory feedback,” *Psychological bulletin*, vol. 60, no. 3, pp. 213–232, May 1964.

[80] M. Wester, Z. Wu, and J. Yamagishi, “Analysis of the Voice Conversion Challenge 2016 evaluation results,” *Proc. INTERSPEECH*, pp. 1637–1641, Sept. 2016.